

Physics fights back

The physical sciences are strongly favoured in President Bush's 2007 budget request — but researchers can't count their chickens yet.

President George W. Bush's State of the Union address on 31 January has been widely reviewed as an uninspiring compendium of small things. But for those directly involved, some of these small things are highly meaningful. For US physicists, the president's public embrace of their work, and his pledge to double funding for it over the next decade, is arguably the best news in the 13 years since the discipline reeled from the cancellation of the Superconducting Supercollider in Texas in 1993.

Bush's unexpected embrace of physics was confirmed in the budget request that he sent to Congress earlier this week. The Department of Energy's Office of Science, which supports the bulk of physics research in the United States, will receive an overall budget increase of 14%, should Congress accept Bush's proposal. Its high-energy physics programme will grow by 8% and its nuclear-physics programme (which funds basic research into atomic nuclei) expands by a spectacular 24%. The National Science Foundation, which supports most non-biomedical research at US universities, also obtains a healthy 8% increase.

These increases are part of Bush's 'American Competitiveness Initiative', which will support more spending on research and education programmes that are deemed relevant to industrial competitiveness. Prompted by public crises at Ford and General Motors, as well as by a general sense of foreboding about China's growing industrial prowess, Congress has been stirring on this issue lately (see *Nature* 439, 517; 2006). It may therefore be generally inclined to support the competitiveness initiative.

Even so, it is by no means certain that physicists will receive even the proposed increases for 2007, never mind the longer-term expansion promised in Bush's speech. This year's proposed cuts in science budgets at NASA demonstrate the danger of relying on bold new initiatives within a general context of fiscal restraint. In NASA's case, the president's plan to send astronauts to the Moon and eventually to Mars is beginning to resemble an albatross around the agency's neck.

It is conceivable that a similar scenario could yet play out at the Department of Energy. In Bush's budget proposal, the increases in science spending are offset by cuts in other programmes at the department, such as the clean-up of former nuclear-weapons plants. These cuts will be fiercely contested by the programmes' respective champions in the Congress.

In this context, advocates for the physical sciences will have their work cut out securing the 2007 spending levels that the president recommends. That said, the competitiveness case for more investment in this arena is a powerful one. As has been noted repeatedly in these pages, inadequate funding at the Office of Science has weakened US physics. It has also distorted the balance of US science spending, driving young people away from scientific work that would be intellectually rewarding, as well as being valuable to industrial competitiveness.

If the budget is approved by Congress, it will give the energy department a long-overdue opportunity to properly use facilities that it has built but can't afford to run. It will also provide more grant opportunities for physical scientists.

But only the long-term investment promised in the president's speech will enable the department to build a new generation of world-class facilities and restore the battered pride of its laboratories. The outlook for this is highly uncertain: the United States must contend with massive financial commitments for the US presence in Iraq and the repair of the hurricane-ravaged Gulf coast. Given the country's financial outlook, US physicists certainly can't bank on their budgets being doubled as the president suggests. Nevertheless, they should be grateful that, after a rocky period, their contribution to the nation is once again being recognized. ■

"The competitiveness case for more investment in this arena is a powerful one."

Network of concern

Only biologists can effectively police the misuse of biological agents.

The array of techniques and compounds that have the potential to be used in bioterror attacks is growing apace, and reaches far beyond a small number of obviously dangerous agents, such as anthrax spores.

At present, the threat is policed mainly by a handful of people in the military and at spy agencies. Interest in the issue has, since the attacks of 11 September 2001, been most acute in the United States.

But bioterrorism is a global threat and, ultimately, it is only vigilance by a much larger network of working biologists that can provide some reassurance in the face of it.

That is the central message of a report released on 31 January by the US Institute of Medicine (IOM). The document, *Globalization, Biosecurity, and the Future of the Life Sciences*, makes a convincing argument that much greater cooperation between scientists around the world will be needed to counter bioterrorism. To this end, it calls on scientists to create a global, grassroots network to discuss and monitor research that might be misused to kill and maim.

Scientists tend instinctively to favour independence and creativity — and to oppose monitoring and regulation. But there are circumstances in which it falls on the community to support



precautionary vigilance over unbridled freedom of action.

The IOM panel, co-chaired by Stanley Lemon of the University of Texas at Galveston, envisages a global biosecurity network not dissimilar to ProMED-mail, a web- and e-mail-based network that helps public-health specialists worldwide share information.

Such a network would enable scientists to exchange views on questions, such as when the risks associated with a particular experiment outweigh its potential benefits. It is extremely difficult to make such calls, as was demonstrated most recently by last year's debate over a paper that modelled a toxin attack on the US milk supply (see *Nature* **435**, 855; 2005). The Lemon report points out that a grass-roots network of scientists interested in bioterrorism issues would at least get a broader section of the community talking and thinking about such issues.

The report also emphasizes the fact that the scope of the bioweapons threat is far wider than commonly imagined. We are entering an era in which scientists will be able to design and build organisms for purposes of their choice, for good or ill. The IOM rightly advises US policy-makers to broaden their consideration of biodefence beyond the 50-or-so 'select agents' that the US health

department has concentrated on until now. That will involve greater consultation with researchers who operate in areas ranging from genetics to nanotechnology, and who do not consider themselves to be involved in biodefence work.

Finally, the IOM calls for more input of up-to-date scientific advice at the US national security agencies, such as the Central Intelligence Agency and the National Security Agency, who are charged with looking out for bioterror-related activity. The work of these agencies is classified, and the report calls for the establishment of a scientific advisory board, with appropriate clearances, to help the agencies interpret the data they collect. This idea was also proposed last March by a commission set up by President George Bush to look at the spy agencies' capabilities in countering weapons of mass destruction. It should be implemented as soon as is practicable.

The containment of potentially dangerous biological knowledge is a formidable challenge — intrinsically even tougher than blocking the proliferation of nuclear weapons. But it is not one that the scientific community can wish away. The IOM panel has performed a valuable service by highlighting some ways in which biologists can step up to it. ■

Japan's research conduct

A framework is required for investigations into scientific misconduct in Japan.

For the past two years, researchers in Japan and elsewhere have been frustrated in their attempts to reproduce a set of RNA interference papers co-authored by Kazunari Taira, a biochemist at the University of Tokyo (see *Nature* **439**, 514; 2006).

The university has now established a committee that will look into Taira's experiments. But some of those involved in the investigation admit that it may not get to the bottom of the issue. The episode points to a lack of preparation for such cases on the part of the university — by far the country's largest and most prestigious research institution. Fault also lies with the government, which has failed to establish a robust framework for investigating alleged scientific misconduct. The lack of such a national framework leaves Japan's universities to look into such allegations independently, and on an ad hoc basis.

The case came to the fore last spring, when the RNA Society of Japan, prompted by members' complaints, requested the University of Tokyo to look into 12 papers produced by Taira's research team. Asked to provide evidence to back up the papers, the team said that it no longer had the notebooks containing the experimental data and that a computer, to which the data had been transferred, had broken down and the data had been erased. The team then struggled to comply with a request that some of its experiments be repeated. Taira has publicly blamed his co-author, Tokyo University biochemist Hiroaki Kawasaki, for mislaying the data. Both men deny any wrongdoing.

Members of a preliminary committee looking into the case complained that they were not given authority to investigate properly. A new committee set up two weeks ago says it will conduct interviews with many researchers on the team. But e-mail exchanges among the researchers will not be used as evidence, and one committee member

admits that, without the notebooks, a determination of whether scientific misconduct occurred is improbable. Committee member Kimihiko Hirao, head the university's school of engineering, says that, unlike the South Korean case of Seoul National University's Woo Suk Hwang, where inside informers pushed the investigation along, they have no direct allegations of misconduct to investigate.

Things could have gone differently. The university did not have a system for accepting and protecting whistleblowers, and therefore could not really expect inside informers to come forward even if misconduct had occurred. There are no clear rules on the penalty if documentation, such as laboratory notebooks or other materials essential to supporting experimental results, is destroyed. And there is no established framework within which such an investigation should proceed.

The University of Tokyo has argued that it lacks the power to investigate the case properly. But two years ago, like Japan's other national universities, it received widespread autonomy from the government. It should use this autonomy to recover from the damage done to its reputation by its unsteady response to this case.

Universities in the United States and Europe have, sometimes with the aid of national misconduct offices such as the US Office of Research Integrity, developed mechanisms allowing them to investigate such cases discretely and comprehensively. A country of Japan's scientific muscle ought to have set up such a body to train universities to investigate fraud and, where appropriate, to administer national sanctions.

Investigations into research misconduct are invariably uncomfortable for all concerned. But the alternative is the pervasive aroma left by cases such as this one, in which the researchers and institutions are neither convicted nor exonerated. It is to be hoped that the investigation now belatedly under way will clear the air. ■

"A country of Japan's scientific muscle ought to have set up a body to train universities to investigate fraud."

RESEARCH HIGHLIGHTS

Stereo smell

Science **311**, 666–670 (2006)

Rats can take an 'olfactory snapshot' of the world in a single sniff, according to research by Upinder Bhalla of the National Centre for Biological Sciences in Bangalore, India, and his colleagues. Their work shows that the rodents can smell in stereo. By contrasting differences between the timing and intensity of smells reaching their two nostrils, the rats determine the direction from which an odour emanates.

The researchers trained thirsty rats (*Rattus norvegicus*, pictured) to drink from a water spout on either their left or right, depending on the direction from which a puff of odour had been delivered. They found that the rat needed just 50 milliseconds to make its decision.



GETTY IMAGES

OPTICS

Longest laser

Phys. Rev. Lett. **96**, 023902 (2006)

A fibre-optic system that could dramatically cut signal power losses in telecommunications has been created by a group led by Sergei Turitsyn and Juan Diego Ania-Castañón at the University of Aston in Birmingham, UK.

The team placed reflectors at each end of an optical fibre and pumped photons into the system. This excited atoms in the fibre and produced another, longer-wavelength set of photons. Because these photons were contained by the reflectors, they acted as a laser amplifier, boosting signals passing through the fibre.

The Aston group showed that there was minimal variation in signal power over 75 kilometres. The device may qualify as the longest laser ever built.

MATHEMATICAL BIOLOGY

Nature by numbers

Bot. J. Linn. Soc. **150**, 3–24 (2006)

Just how mathematical is nature? Answers to this question invariably invoke the idea that the optimal arrangement of leaves or petals in a plant (phyllotaxis) is related to the Fibonacci series, generating groupings of 2, 3, 5, 8 and so on.

But after a critical look at the data, Todd Cooke of the University of Maryland, College Park, concludes that the example may owe more to 'Pythagorean mysticism' than to hard evidence. The spiral patterns in structures

such as pine cones and sunflower seed heads do seem to relate to the Fibonacci sequence, but contrary to common belief, they don't correspond to optimal packing. Also, increases in the number of leaves on a growing stem do not follow the Fibonacci sequence, as would be expected if it was fundamental to phyllotaxis.

QUANTUM PHYSICS

Everything is illuminated

Phys. Rev. Lett. **96**, 037402 (2006)

In today's computers, information is encoded in units known as 'bits'. Each bit can take one of two possible electronic states, '0' or '1', switched by a voltage.

The potential of a quantum computer, however, depends on its quantum bits, or 'qubits', sometimes being in both the 0 and 1 states at once. Now, Stefan Stufler of the University of Paderborn, Germany, and his

co-workers show how the much more complex state of a qubit can also be set by a simple voltage.

The team used the absence and presence of charges in a speck of semiconductor, known as a quantum dot, to represent a qubit in states 0 and 1, respectively. Charges were created by illuminating the dot with two short laser pulses. The researchers showed that the voltage applied across the dot controlled the mixture of states in the qubit.

STRUCTURAL BIOLOGY

Pulling power

Proc. Natl Acad. Sci. USA **103**, 1244–1247 (2006)

For proteins whose structures resist analysis by conventional means, such as crystallography or computer simulation, Hendrik Dietz and Matthias Rief of the Technical University of Munich, Germany, introduce a brute-force approach.

They inserted chemical groups that can be crosslinked at selected locations in a protein, and used these to connect several of the proteins into long chains. When the chain was pulled taut with the tip of an atomic-force microscope, the portions of protein between the crosslinks unfolded one by one, causing the length of the polypeptide chain to increase in steps. From measurements of the change in length, the initial separation of the crosslinking groups in the folded protein could be calculated. Repeating this for different pairings of three crosslinking groups allowed the relative positions of the inserted groups to be determined by



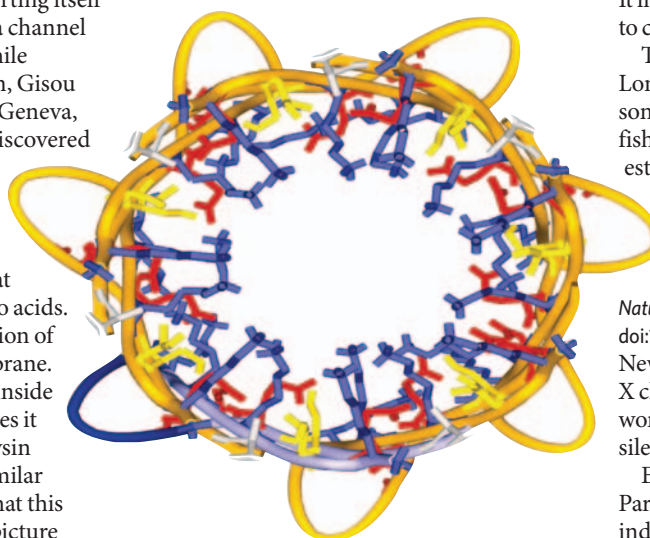
triangulation. Measurements made on the structure of green fluorescent protein agreed closely with crystallographic data.

CELL BIOLOGY

Riveting toxin

EMBO J. doi:10.1038/sj.emboj.7600959 (2006)
Like many toxins produced by bacteria, aerolysin causes cell death by inserting itself into the cell membrane, creating a channel through which vital ions leak. While studying the insertion of aerolysin, Gisou van der Goot of the University of Geneva, Switzerland, and her colleagues discovered that this toxin uses a novel mechanism to anchor itself in the membrane.

The team identified a short stretch in the aerolysin protein that comprises five hydrophobic amino acids. These form a tip that drives insertion of the toxin into the cell's lipid membrane. Once the tip reaches the aqueous inside of the cell, its hydrophobicity causes it to turn back on itself, fixing aerolysin into the membrane like a rivet. Similar stretches in other toxins suggest that this mechanism may be general. The picture (right) is a top-down view of seven aerolysin proteins assembled into one channel.



ASTRONOMY

A wee wobble

Astron. Astrophys. **447**, 361–367 (2006).
The search for planets outside our Solar System was recently boosted by a relatively new method based on the gravitational lensing of light by a star–planet system. The technique found a planet about 5.5 times the

mass of Earth (*Nature* **439**, 437–440; 2006).

But there's still plenty of life in the more established Doppler technique, which finds planets by measuring the wobble that they induce in the orbit of their parent star.

Stephane Udry and his colleagues at the High Accuracy Radial velocity Planetary Search (HARPS) project at La Silla, Chile, report that they have found a planet with a

mass 14 times that of Earth. The planet orbits its star once every 15.6 days at a distance of just 17 million kilometres.

ECOLOGY

Plenty more fish in the sea

Science **311**, 660–663 (2006)
Vast shoals of fish have been monitored in real time across many tens of kilometres using a novel remote-sensing technique.

The technique, developed by Nicholas Makris of Massachusetts Institute of Technology in Cambridge and his colleagues, uses acoustic reflections to detect the position and population density of shoals of fish. The new technique sees much further than conventional fish-finding sonar because the researchers use a trick to reduce the rate at which the sound intensity drops with distance. It involves using the sea surface and sea floor to channel low-frequency sound waves.

The researchers tested the method off Long Island in New York State, detecting some shoals containing tens of millions of fish. The technology should help to improve estimates of fish populations, says Makris.

GENETICS

X marks the spot

Nature Cell Biol. doi:10.1038/ncb1365 (2006); *Science* doi:10.1126/science.1122984 (2006)

New observations hint at how women's two X chromosomes may exchange signals in the womb before one of the chromosomes is silenced forever.

Edith Heard of the Curie Institute in Paris, France, and her collaborators and, independently, a group led by Jeannie Lee at the Howard Hughes Medical Institute in Boston, Massachusetts, tracked the movement of X chromosomes in embryonic mouse cells using fluorescent labelling. They saw that regions of the chromosomes known to be involved in X inactivation drew close to each other during the first few days of development. The researchers say this could allow signals to be passed between the chromosomes. When parts of these regions were deleted, the pairing did not happen and X inactivation was disrupted.

JOURNAL CLUB

Sun Kwok
University of Hong Kong, China

An astronomer is bugged by the scarcity of one of life's vital elements in space.

The idea that we are a product of the stars dates back to the 1950s. Then, astronomers realized that most chemical elements in the Universe are made by nuclear reactions in the stars.

Unsurprisingly, the elements that are most common in living things, including oxygen, carbon

and nitrogen, are also among the most cosmically abundant. But there is one fascinating exception.

Phosphorus is important in biochemistry: it is involved in energy transfer and membrane structure, for example. The relative abundance of phosphorus in the human body is several orders of magnitude greater than in the Solar System, where it is only the seventeenth most common element.

So, how did phosphorus concentrate on Earth, ultimately becoming part of us?

This conundrum is discussed by Enrique Maciá of the Complutense University of Madrid, Spain (E. Maciá *Chem. Soc. Rev.* **34**, 691–701; 2005). He reviews the synthesis of phosphorus and measurements of its abundance, and speculates that the element preferentially condenses into solids, becoming incorporated into comets that deliver phosphorus to Earth.

That provides a mechanism for the element's arrival, but what about the forms in which it floats round space? We have detected many molecules containing carbon,

nitrogen and oxygen in the interstellar medium and around stars. The only known phosphorus-bearing astrophysical molecules are phosphorus mononitride and carbon monophosphide.

Motivated by this apparent paradox, I am, with others, searching for other phosphorus-containing molecules using the satellite Odin and ground-based telescopes. We hope that such detections will shed light on phosphorus's journey from the stars to Earth.

NEWS

Dealer unearths Hooke's Royal Society notes

The battered folio of notes looks like many others that have landed on the desk of Felix Pryor, manuscript expert at Bonhams auction house in London. But when he leafed through the pages of intricate seventeenth-century handwriting, Pryor quickly realized he was looking at the most exciting scientific documents he had ever handled: a set of original minutes and commentary that detail the birth of modern science. "My jaw hit the floor," he says.

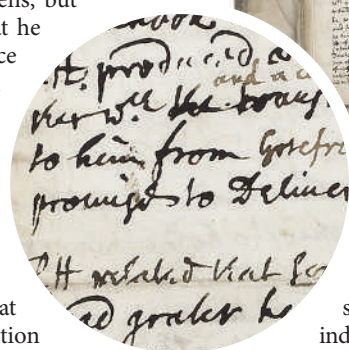
The manuscripts are due to go on sale next month with an estimated value of £1 million (US\$1.8 million). They capture the early days of Britain's oldest research institution, the Royal Society, through the eyes of Robert Hooke — a brilliant physicist, chemist and mechanic who during his career engaged in well-documented rows with contemporaries such as Isaac Newton. Historians who have seen the papers say they settle a major controversy in the development of timekeeping and provide fascinating insights into the fledgling UK research body.

Most of the notes date from Hooke's tenure as secretary of the society between 1677 and 1682, a time when Newton was developing his theory of gravity and Gottfried Leibniz was working on calculus. Transcriptions of Hooke's minutes, which record the group's weekly Thursday meetings, were printed at the time and survive to this day. But the discovery of the

originals fills in several gaps created when the copies were made.

A second section sheds light on a 300-year-old quarrel over the invention of the spring-balance watch, a major advance in timekeeping. The invention is often attributed to the Dutch physicist Christiaan Huygens, but Hooke always claimed that he had demonstrated the device at a Royal Society meeting in June 1670, five years before Huygens patented his. As the society's transcribed minutes make no mention of such a demonstration, many historians have questioned Hooke's claim.

The notes may change that thinking. The second section was written by Hooke as he looked through original minutes handwritten by his predecessor as secretary, Henry Oldenburg. Hooke had removed a sheet of Oldenburg's minutes and placed it in his folio. It contains details of the pocket-watch experiment that Oldenburg never transcribed into the official records. "Hooke must have been incensed when he found it," says Lisa Jardine, a historian at Queen Mary, University of London, and a biographer of Hooke.



Science historians are desperate for Hooke's notes to remain accessible to study.

BONHAMS

Other parts of the manuscript suggest that Hooke was indeed furious. Oldenburg's printed minutes for 1676 omit another of Hooke's experiments, this time on magnetism. In his commentary, Hooke notes: "The dog has entered nothing but Left a blank."

The documents were found last September in a cupboard in a house in Hampshire, southern England, during a routine valuation of other items. The owners say they have no idea how they came by the manuscript, and that it had been in the cupboard for as long as they can remember.

Drop in HIV infection rate used to justify focus on morality

A study claiming good news on Africa's fight against AIDS has reignited conflicts over the effectiveness of HIV prevention campaigns, particularly the merits of focusing on abstinence and monogamy.

Researchers reported last week that HIV infection rates in Zimbabwe fell between 1998 and 2003 from 23% to 20.5% (S. Gregson *et al. Science* 311, 664–666; 2006). They claimed that the gains came thanks to behavioural changes by Zimbabweans, who were having sex at a later age and with fewer partners. Women also reported more frequent condom use.

Some scientists and government officials hailed the finding as showing that campaigns persuading people to change their behaviour can succeed. "This supports the idea that



Greater condom use and monogamy are being linked to a fall in HIV in Zimbabwe.

behavioural change has resulted in a reduction in new infections in this study population," says Richard Hayes, an epidemiologist at the London School of Hygiene and Tropical Medicine.

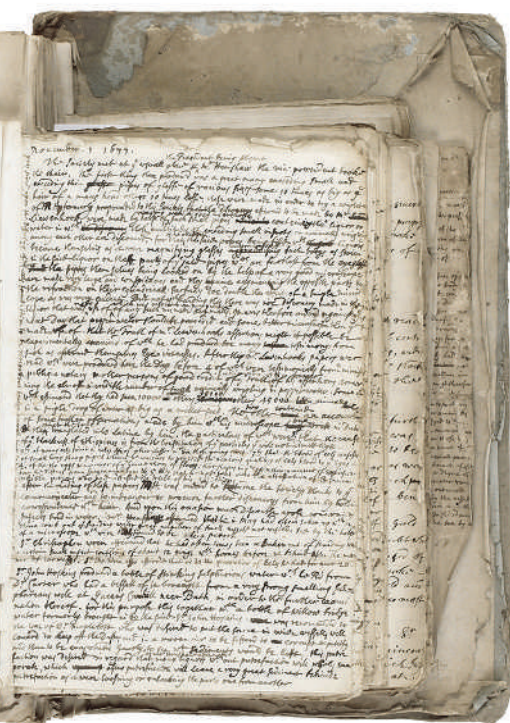
More controversially, US officials claim that the study validates their foreign policy. A third

of US money spent on AIDS prevention overseas goes on programmes promoting abstinence and being faithful to one partner — the 'A' and 'B' in the 'ABC model', where 'C' stands for condoms. Critics say that focus is based on morality, not science, and can be unrealistic, especially for women who may have no choice about with whom they or their partners sleep.

Deputy US global AIDS coordinator Mark Dybul says the Zimbabwe study and earlier gains in Uganda help to put such criticisms to rest. "The greatest behaviour change was in abstinence and fidelity," he points out.

But Simon Gregson of Imperial College London, who led the Zimbabwe study, insists that "we need to be promoting all the different prevention possibilities".

R. SMITH/CORBIS



Jardine and others say it is vital that the manuscript remains accessible after its sale, so that it can be properly studied by science historians. But there is no guarantee that will happen. The Royal Society says it wants to bid for the documents, but lacks the funds. "I believe there is a lot of other serious stuff in there," says Jardine. "I'm desperate that it goes to an accessible home."

Should a foreign collector triumph in the auction, the UK government is likely to impose a three-month delay on issuing an export licence for the manuscript. Such a move would be designed to enable British museums to raise funds to mount a rival bid; UK law dictates that a museum could take possession if it matched the winning bid. ■

Jim Giles

And some warn that other factors have been overlooked. Maria Wawer of Columbia University in New York, co-leader of a long-running study in Uganda, thinks deaths from AIDS play a role. Although prevention programmes deserve some credit, she says, "there's no reason to believe mortality is not contributing to this". After a decline in the late 1990s, HIV prevalence in Uganda now seems to be levelling off or even climbing back up.

The drop in Zimbabwe could also have been affected by violence and unrest, which have escalated there since 2000. Epidemiologist Christopher Beyrer of Johns Hopkins University in Baltimore, Maryland, says forced relocations and the departure of millions of young people seeking work may have skewed the study's results. ■

Erika Check



INTELLIGENT-DESIGN FILM SCREENING

Read our report of the first night of *Flock of Dodos: the Evolution-Intelligent Design Circus* in Kansas City

www.nature.com/news

Possible planets left with no name

Astronomers trying to define a 'planet' won't announce a decision until August at the earliest, and may duck the issue entirely. In the meantime, researchers are left unable to name some newly discovered objects.

The issue was raised when astronomers led by the California Institute of Technology's Mike Brown reported finding objects that came close to Pluto in size, including Quaoar in 2002 and Sedna in 2004. It came to a head in July 2005, when the group announced the discovery of an even larger object, temporarily dubbed 2003 UB₃₁₃. Last week, scientists confirmed that this is substantially bigger than Pluto (F. Bertoldi *et al.* *Nature* 439, 563–564; 2006).

Until astronomers decide what counts as a planet, UB₃₁₃ cannot be officially named. The International Astronomical Union (IAU) catalogues such objects, but different committees are responsible for naming major planets, and mere chunks of rock or ice. The indecision "really interferes

with our work", says Brian Marsden, head of the Minor Planets Centre in Cambridge, Massachusetts, which catalogues thousands of objects every year. Brown's discoveries, 2005 FY9 and 2003 EL61, are also waiting to be named.

The IAU set up a 19-strong committee in 2004 to resolve the issue, which reported to the executive committee in November. Some experts suggested that anything spherical that orbits the Sun and is more than 2,000 kilometres across should be called a planet. That size limit would include Pluto and UB₃₁₃, and was selected as "arbitrary but sensible", according to Iwan Williams, president of IAU's planetary-systems sciences division.

But other suggestions included stripping Pluto of its title, because thousands of objects occupy the same area of space, known as the Kuiper belt. "We should speak of eight major planets," says Marsden. "People say schoolchildren will be upset, but so what? It's our job to educate them."

Others disagree. "Pluto is obviously a planet," says Alan Stern of the Southwest Research Institute in Boulder, Colorado, who heads the New Horizons mission that has just set off to Pluto. He argues that anything big enough to form a sphere as it orbits a star should be called a planet — including the asteroid Ceres. This list could be broken down into categories such as 'terrestrial planet' or 'gas giant' (see *Nature* 437, 456–457; 2005).

The matter is now likely to be referred to the entire IAU membership at a meeting in Prague this August. Robert Williams, who sits on the executive committee, isn't optimistic about a resolution: "Nineteen experts wrestled with this for six months and didn't reach a conclusion." He favours waiting until more is known about the edges of our Solar System, and the planets around other stars. "My recommendation is we're not ready to move on this yet." ■

Mark Peplow

NASA-JPL



New arrival: UB₃₁₃ is bigger than Pluto but hasn't yet been formally named because of indecision about its status.

ON THE RECORD

“Old octopuses become what we call senescent, or senile... and sometimes their actions are very inappropriate.”

Jim Cosgrove of the Royal British Columbia Museum speculates about why an octopus recently attacked a small research submarine.

“Please note that from now on, Orion's Belt will be replaced by Chantelle's Thong.”

Columnist Lucy Mangan derides the decision to replace London Planetarium programmes with a show about celebrities.

“You cannot ethically... infect humans and see if they get fat.”

Leah Whigham of the University of Wisconsin, Madison, explains why she infected chickens, not people, with a virus linked to obesity.

Sources: CBC, The Guardian, The Telegraph

SCORECARD**Moon mining**

Russia's state-owned space company says that the nation plans to establish a Moon base and mine lunar helium-3 as a fuel for fusion reactors.

**Placebos**

Sham acupuncture works better than a cornstarch pill at easing arm pain, says a comparison published in the *BMJ*.

**Telegrams**

After 150 years, Western Union has sent its last telegram. The company, which also does money orders, says it will now focus on “financial services”.

OVERHYPED**Bird flu in Austria**

Panic spread throughout Vienna last week after citizens reported some 40 dead songbirds. Autopsies showed that the H5N1 flu virus was not to blame. The larks, it seems, had developed an appetite for fermented berries, and got so drunk that they flew into windows.

One place, one parent, two species

Evolutionists have provided what they hope is a definitive answer to the question of whether ‘sympatric speciation’ can occur in the natural world. This is where two or more species diverge from a common ancestor despite sharing the same habitat. The process has been taught as part of evolutionary theory for decades, but has been notoriously difficult to prove.

There are candidate cases, such as where parasitic species seem to have diverged on the same host. But solid examples are hard to pin down, as they require showing that two or more species descended from a single ancestor, and that the lineages did not spend any time apart during the entire course of their divergence — often a span of several million years.

Two studies in disparate parts of the globe offer what their authors say are the best shots so far at such a proof. On page 719 of this issue, Axel Meyer and colleagues at the University of Konstanz in Germany describe two fish species that have apparently diverged despite sharing a relatively small lake (5 kilometres across) in Nicaragua. According to Meyer and his team, it's next to impossible that two such closely related species could have ended up in this very young (less than 23,000-year-old),

Evidence found: two species of palm on an island in Australia seem to have diverged side by side.

isolated lake through two colonization events.

In a second study, Vincent Savolainen of Britain's Royal Botanic Gardens in Kew and his colleagues report that two species of palm tree seem to have diverged on a remote Australian island (V. Savolainen *et al. Nature* doi:10.1038/nature04566; 2006). In this case, the remoteness of the island and the shared genetic traits of the species make it extremely unlikely that the trees are descendants of two ancestors that arrived at the island on separate occasions.

“Almost no proposed example of sympatric

Political strife set to delay EU funds

MUNICH

A power struggle between the European Parliament and member states of the European Union (EU) threatens to delay the start date of the EU's next funding programme for research — and that of its new flagship funding agency, the European Research Council (ERC).

In December, the EU's 25 heads of government agreed a total budget for 2007–13, including a 75% increase in research spending by 2013. If the money rises steadily over that period, the budget for the seventh EU Research Framework Programme (FP7) will be roughly €50 billion (US\$60 billion). That's €20 billion

less than originally proposed by the European Commission, which aimed to double research spending by 2013.

A research-friendly majority in the European Parliament had also hoped for more. On 18 January, the parliament rejected the proposed EU budget by 541 votes to 56, forcing governments to renegotiate. “We don't reject the budget lock, stock and barrel,” says Jorgo Chatzimarkakis, a liberal democrat MEP from Germany. “But we can't accept cuts in research that are disproportionately larger than in agriculture and structural funds.”

Although the possibility of an extra few billion euros for research may please

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

MARK A. JOHNSON/CORBIS

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

speciation can be considered absolutely beyond doubt,” comments Jerry Coyne, an evolutionary biologist at the University of Chicago, Illinois. But in both of the newly reported cases, he says, sympatric speciation seems the most likely explanation. William Baker, part of the Kew team, agrees: “My first reaction is that it’s hard to imagine a case being neater than ours — it sounds immodest but it’s almost the ideal scenario.”

The big question now is whether sympatric speciation is widespread or rare. Although it is relatively easy to make a convincing case for a remote island or small isolated lake, things will be much tougher when biologists turn

to more complex ecosystems such as East Africa’s great lakes, home to hundreds of related fish species. “What would overturn our view of evolution would be to find that a fairly substantial proportion of all speciation events occur in sympatry,” Coyne says.

The concept that such speciation is widespread was largely resisted by the great evolutionist Ernst Mayr, who died last year, says Meyer. Such was Mayr’s influence among biologists that he engendered a widespread bias against the idea. “But I think sympatric speciation will turn out to be more common than people thought,” Meyer adds. ■

Michael Hopkin

Europe’s science community, research managers fear continued political wrangling will lead to delays. Governments are unlikely to substantially revise the overall budget, which they agreed, after prolonged negotiations, just before the British EU presidency ended. So it is up to the commission to deliver a revised proposal for FP7. But insiders say departments are fighting over how to cope with the reduced budget.

“The calendar for the adoption of FP7 is now extremely tight,” Janez Potočnik, the EU’s research commissioner, told a parliament committee on 26 January. “If we are to deliver a Framework programme in time for 2007...we must ensure that there are no delays on either side.”

“We’re anxious to get started in January.”

The commission had earmarked €1.5 billion per year for the ERC, which will fund basic research, but that figure is now unlikely. “€1 billion per year is the absolute minimum required for the ERC to make sense,” says Carl-Henrik Heldin, director of the Ludwig Institute of Cancer Research in Uppsala, Sweden. “We’re anxious to get started in January, and get the first money paid out by the end of 2007.”

Peter Dröll, head of Potočnik’s staff, says FP7 could make its schedule, provided the EU budget is approved by April. But that’s far from certain. “I reckon we’ll need at least until the summer break,” says Angelika Niebler, a Christian democrat MEP who acts as rapporteur for the ERC. ■
Quirin Schiermeier

Space rocks wanted: cash paid

SAN DIEGO

Meteorites have become prime targets for high-end curio dealers, reaching as much as US\$10,000 a gram at auction. Professional scavengers have reacted with a flood of space rocks, forcing scientists to take radical steps and set up a centre at the University of Arizona that will buy samples on the open market.

"We don't want to shut down the trade in meteorites, because we can't," says cosmo-chemist Dante Lauretta of the university's Lunar and Planetary Laboratory in Tucson. But the aim is to preserve as many of the finds as possible for scientific analysis, he says.

Meteorites made of material ejected from the Moon and Mars command the highest prices. But their popularity is pushing up the cost of all meteorites — including those known as chondrites, which are highly prized by scientists because they formed early enough to offer insight into the origins of the Solar System.

Over the past decade or so, the promise of bounty has spurred meteorite hunters to scour areas such as the Sahara Desert, the Gobi Desert and Patagonia, where arid conditions preserve specimens for millions of years.

Historically, scientists at a number of insti-



Pebble dash: collectors race to pay stratospheric sums for meteorites like this.

tutions, including the University of Arizona, have analysed material for collectors or dealers — in return receiving a portion of the meteorite for scientific study. "But there are so many meteorites coming out of the Sahara now that scientific institutions are overwhelmed," says Lauretta. "It would take us two years to analyse everything we have."

So he and his colleagues have set up the Southwest Meteorite Center to buy samples of meteorites before they are divided up and disappear unrecorded into private collections.

With an initial fund of about \$200,000, the

centre will buy meteorites wherever it can, as well as continuing to analyse rocks for dealers in return for samples. The researchers began the search for samples on 4 February at a meteorite auction held ahead of the annual Tucson Gem and Mineral Show, where the world's major players wheel and deal. "It was quite an adventure," said Lauretta, after appealing to a packed house of about 250 dealers and well-heeled customers. "We were very well received."

The top sale of the day was \$6,750 for a newly discovered 11-kilogram piece of the Brenham pallasite. This famous meteorite, which consists of olivine crystals in an iron-nickel matrix, struck

Kansas in 1882.

The centre will also buy collections from private individuals, and is raising funds to create an endowment for future purchases. The samples will be stored in a climate-controlled facility and made available to scientists who want to study them.

Long-time meteorite dealer Marvin Killgore of Payson, Arizona, will curate the repository. He has lent it a significant portion of his personal collection, which consists of 3,300 kg of meteorites from 37 countries.

Rex Dalton

E. ENGELER/AP

Journal lays bare remarks from peer reviewers

Editors of a journal launched this week are out to revolutionize peer review. By publishing signed reviews alongside papers, they hope to make the process more transparent and improve the quality of the articles. But although journal editors seem intrigued by the experiment, most say they'll take some persuading to change the traditional, anonymous system.

At *Biology Direct*, an open-access journal launched by BioMed Central on 6 February, manuscript editors and peer reviewers will, in effect, be merged into one editorial board. Prospective authors will approach board members and if three agree to review a paper, it will be accepted.

Reviewers' comments will be signed and published with the final paper, along with responses by the author. An author has the right not to make suggested changes, but the suggestions will be there for anyone to see. An author who disagrees with the comments can retract the paper.

Several journals, including the *BMJ* (British Medical Journal) and the *Medical Journal of Australia*, have experimented with naming peer reviewers, but *Biology Direct* is going further by routinely posting those reviews as part of the paper.

"I like the direct relationship the author can have with a reviewer, and the transparency of the end result," says David Lipman, director of the US

National Center for Biotechnology Information and one of the journal's lead editors. He believes readers will get a more nuanced picture of science.

"We don't have that artificial, black-and-white situation where, because it got through peer review, it is all fine," Lipman says. "It will be those interactions with the peer reviewers that make it interesting."

Responses to the idea have been positive. "I love the fact that David Lipman is doing this," says Drummond Rennie, deputy editor of the *Journal of the American Medical Association*. "I have always felt that the only ethically sound system of review was one where everyone knew everyone's identities."

Diane Sullenberger, executive editor of the *Proceedings of the National Academy of Sciences*, says she'll be watching *Biology Direct* with interest, although her journal will not open its peer-review system any time soon. "If there was evidence in the literature that open peer review really had a significant advantage over blind review I think we'd see more of it," she says.

Lipman says that the new journal's policies are likely to evolve. "It is an experiment," he says. "But I think the overall approach can't help but succeed. If we are really successful, the better journals will take what they think works from it."

Emma Maris

US targets roles of genes and environment in disease

The US National Institutes of Health was this week set to announce two ambitious projects in genetics.

The Genes and Environment Initiative and the Genetic Association Information Network (GAIN) are both aimed at identifying the genetic and environmental contributions to common, complex diseases.

The projects will collect genetic and environmental information — such as diet and exposure to chemicals — from thousands of patients with specific diseases and from an equal number of controls. Researchers hope that this will allow them to pinpoint factors that contribute to conditions such as heart disease and Alzheimer's.

The Genes and Environment Initiative will be wholly funded by the government, and officials in the Bush administration have asked Congress for \$40 million for the project in their 2007 budget request (see page 644). In total the NIH will seek \$160 million to fund the initiative over four years.

GAIN, meanwhile, will be funded in part by private companies. The drug firm Pfizer, for example, has so far pledged \$20 million, and biotech company Affymetrix will pay for studies of at least two diseases, which will cost \$3 million to \$6 million each. The programme is scheduled to begin this summer, once peer reviewers have selected the first diseases for investigation.

Boston biosafety lab gets all-clear for construction

After years of protests, Boston University's high-level biosafety lab finally secured federal approval last week.

The lab will be one of two National Emerging Infectious Diseases Laboratories, and will include biosafety level 4 facilities. The \$178-million building will be located at the university's medical centre in south Boston.

The biosafety 4 lab will be used for



Protesters gather to vent their feelings about Boston University's proposed biosafety lab.

Dwarf galaxies yield some of dark matter's secrets

Astronomers have come up with the most precise description of dark matter yet, after measuring the temperature of the invisible material holding nearby dwarf galaxies together.

Although dark matter cannot be seen, its existence can be inferred from its gravitational interactions with the stars around it. Gerry Gilmore of the University of Cambridge, UK, and his team used the Very Large Telescope array at the Paranal Observatory in Chile to find that 12 dwarf galaxies (such as the one pictured) orbiting the Milky Way each contain the same amount of dark matter — roughly 30 million times the mass of the Sun. The team concludes that this is the minimum amount of dark matter needed for a stable clump to hold together.

This implies dark matter isn't as cold as theorists had thought. Whatever the particles are made of, they must be moving at about 9 kilometres per second, equivalent to a bulk



J. HAMBSCH & R. GENDLER

temperature of around 10,000 °C. "These are the first properties of dark matter, other than its existence, that we've been able to determine," says Gilmore.

research on highly infectious pathogens such as Ebola virus. When plans for the facility were announced back in 2003, they sparked protests from local scientists and residents (see *Nature* 428, 785; 2004). A state environmental review is still pending, but this is not expected to halt the project.

The university says that construction should begin later this month and is scheduled to take 30 months to complete.

Fall-out over faked clones continues to shake Korea

A truck driver torched himself to death in Seoul last week after distributing leaflets calling for disgraced cloning expert Woo Suk Hwang to continue his research. The event was symbolic of a scandal that continues to shake South Korea.

On 31 January, two bioethicists retracted a paper in which they outlined the informed-consent procedures for egg donation that they had devised in collaboration with Hwang (K. W. Jung and I. Hyun *Am. J. Bioeth.* 6, W19–W22; 2006).

And on 2 February, Korea's National Bioethics Committee confirmed rumours about the way Hwang procured eggs. It found that Hwang had forced junior members of his lab to donate eggs, and that he used more than 2,221 eggs in his research rather than the 400 or so he acknowledged using. Thirty-five women's groups plan to sue the government for supporting Hwang's research and neglecting issues related to egg procurement. About one-fifth of the donors, many of whom weren't told of the risks, are now suffering side effects.

Hwang is being investigated over a variety of legal matters, including misuse of state funds, fraud in applying for funding based on knowingly faked results, and breaching the country's bioethics law.

NASA press office blasted for overstepping the mark

NASA's public-affairs machine came under fire last week for allegedly clamping down on climate researcher James Hansen, of the agency's Goddard Institute for Space Studies in New York. He claimed he was prevented from speaking to reporters after he called for action to dramatically reduce greenhouse-gas emissions.

Most of the focus was on George Deutsch, a political appointee in NASA's press office in Washington, who, according to *The New York Times*, refused a reporter's request for an interview with Hansen (see www.nature.com/news/2006/060130/full/060130-11.html). In a separate incident reported by the same newspaper, Deutsch insisted that the word 'theory' accompany each reference to the Big Bang on a NASA website. "It is a religious issue," he said, as the Big Bang would discount intelligent design by a creator.

The news brought a swift response from the chairman of the House Committee on Science, Sherwood Boehlert (Republican, New York), who ordered his staff to look into the allegations. NASA administrator Mike Griffin issued his own statement, saying: "It is not the job of public affairs officers to alter, filter or adjust engineering or scientific material produced by NASA's technical staff."

NEWS INTERNSHIP

Nature is seeking an intern reporter to work full-time in either its London or Washington office from July 2006 for six months. Applicants should be self-starting and have a keen news sense. This is a paid position. The intern will write news and news features for *Nature* as well as online news for *Nature's* website news@nature.com. Please e-mail a cover letter, résumé and three clips to Alexandra Witze, senior news and features editor (a.witze@nature.com) by 15 February. Put 'internship application' in the subject line.

SPECIAL REPORT

Physics wins the US budget race

Energy and competitiveness are the buzzwords in George W. Bush's proposal for spending in the next fiscal year. But will the president's push really benefit researchers?

On 6 February, physical scientists in the United States got some rare news: a big funding boost for next year, courtesy of President George W. Bush. But it seems that the budget for other disciplines must remain flat, or drop slightly, to make up for the physics push in a science budget that drops slightly overall.

As part of a \$5.9-billion 'American Competitiveness Initiative', the president has proposed a funding increase of almost a billion dollars for research at the National Science Foundation (NSF), the Department of Energy (DOE) and the National Institute of Standards and Technology. Together, these bodies finance the lion's share of physical-sciences research in the United States. An additional \$380 million will be put towards applied energy research, in an effort to reduce the nation's dependence on foreign oil.

"This funding will support the work of America's most creative minds as they explore promising areas such as nanotechnology, supercomputing and alternative energy sources," Bush said on 31 January during his annual State of the Union address.

But science budgets at other agencies, such as the National Institutes of Health and NASA,



Bush has come through for physicists, but the budget is harsher on others.

are not likely to be so generous (see 'Biomedical funding remains flat' and 'Science at NASA takes a hit'). If Bush's proposal is accepted by Congress, the overall budget for 'federal spending in science and technology' — a measure formulated by the National Academies — will drop by 1% from last year to \$59.8 billion (see chart). "The point is, we're prioritizing," says John Marburger, the president's science adviser.

In an annual ritual, Monday saw the US president release his suggestions for spending government money: a \$2.8-trillion plan for fiscal year 2007, which begins on 1 October 2006. The budget includes cuts in health care,

education and the environment, and increases for defence and homeland security. It does not include money to cover the war in Iraq or the aftermath of Hurricane Katrina, both of which have contributed to a sense that funds are particularly tight this year.

In science, Bush proposes that the largest percentage increase in funding should go to the DOE, which oversees most of the nation's large physical-science facilities. Under his plan, the department's Office of Science will see an impressive 14% increase, to

\$4.1 billion, in 2007. Among specific projects, an extra \$100 million will go to the Spallation Neutron Source, which is set to begin operations this year at Oak Ridge National Laboratory in Tennessee. An additional \$87 million will go to nuclear-physics facilities, including those at Brookhaven National Laboratory in New York and the Thomas Jefferson National Accelerator Facility in Virginia; this is a relief after earlier budget cuts had raised fears of cutbacks or even closures at these labs. And \$60 million will go to funding the US commitment to ITER, an international fusion experiment based in Cadarache, France.

The energy department also wants to spend

Biomedical funding remains flat

Spending on biodefence and pandemic influenza mark the few bright spots in an otherwise dreary budget proposal for the National Institutes of Health (NIH). President George W. Bush has asked Congress to give the agency \$28.6 billion in 2007 — the same amount it received in 2006, when it absorbed its first spending cut in 36 years (see *Nature* **439**, 129; 2006).

All of the NIH's major institutes and centres will see their budgets fall under the 2007 plan, except for the National Institute of Allergy and Infectious Diseases and the

NIH director's office (see briefs, opposite). But \$17 million is slated for preparations for the possibility of pandemic flu. The agency is also launching a programme to study the interactions between genes and the environment, and their effects on disease (see page 643).

The NIH's roadmap for medical research, which lays out a long-term plan to translate basic research findings into clinical applications, will also get a \$113 million increase, rising to \$443 million. It is one of NIH director Elias Zerhouni's hallmark initiatives, although the boost is

less than he asked for when he started the project in 2003 (see *Nature* **425**, 438; 2003).

Overall, the NIH estimates that the budget constraints will lead to a 1.7% drop in the number of grants it awards. Nevertheless, Zerhouni vows that the number of competing grants, which include those awarded to new projects, will not be cut. "I think it's very important that new investigators do not get discouraged," he says.

If Congress agrees to the president's request, it will be the fourth year in a row that the NIH's budget has not kept pace with the

rising cost of research due to inflation. As a result, says David Moore of the Association of American Medical Colleges, the agency's budget could end up being 10% less than it was at the end of 2003, in inflation-adjusted dollars. "We're obviously very concerned," Moore says.

He and other biomedical research advocates plan to ramp up their lobbying on Capitol Hill, as some lawmakers feel that the NIH should be content with the five-year doubling of its budget, which was completed in 2003.

Erika Check

P. MARTINEZ MONSIVALS/AP



COULD A SPRINKLING OF DIRT SAVE THE GLACIERS?

Knowing how spiky glaciers form could give clues about how to slow ice melt.

www.nature.com/news

Science at NASA takes a hit

Astronomy and planetary science would suffer under the Bush budget, which plans to send astronauts back to the Moon and keep flying the space shuttle to the International Space Station.

Under the president's \$16.8-billion request for NASA, the agency's overall budget would stay essentially flat, rising only 1% from last year.

But the Moon exploration programme would grow 30%, to \$3.98 billion in 2007. And the space shuttle and space station would get an additional \$2.6 billion between 2007 and 2010 over what the agency projected last year.

To balance those increases, 'science' spending would be held to a 1.5% increase next year and 1% in following years. Standing at \$5.25 billion today, the science budget was last year projected to climb to \$6.8 billion by 2010. But NASA administrator Michael Griffin now says: "We cannot afford such growth."

The list of casualties includes projects that are cancelled or deferred indefinitely, such as the Terrestrial Planet Finder and a Mars sample-return mission. Others have just been delayed, such as the Space Interferometry Mission, which will slip three years.

Some, including the SOFIA airborne observatory, are in limbo pending a budgetary review (see *Nature* **439**, 515; 2006). Still others that had hoped for start-up funding in 2007 will now have to wait, including a mission to Jupiter's moon Europa that Congress had specifically requested. And research grants to scientists will be reduced, particularly in the area of astrobiology.

The complaints were loud and immediate. "They're starving science to feed an emaciated space shuttle," fumes a Democratic congressional staffer.

Tony Reichhardt

BUDGET BITES

Giving credit

By far the largest part of the president's new 'competitiveness initiative' would go to private industry, in the form of a \$4.6-billion tax credit. The credit is currently awarded by Congress to companies that invest in research and development on a year-to-year basis, but the administration argues that this makes it hard for industries to plan investment in long-term research. By making the credit permanent each year, the White House hopes to encourage growth in the nation's approximately \$200-billion industrial-research sector. But it won't come cheap: estimates show the credit will cost federal coffers \$86.4 billion by 2016.

Down at defence

Although the president's plan would increase physical-sciences spending in the civil realm, the funding of basic and applied research in the defence department would drop by 11% to \$5.9 billion. The cuts are partly an attempt to scale down the department's huge budget, but they also reflect a shift in priorities, away from fundamental science towards weapons development.

Biodefence boost

Meanwhile, spending on biodefence continues to rise. The president has asked for an extra \$178 million compared with last year in biodefence spending for the Department of Health and Human Services. This includes a stash of \$160 million that would help specific projects to compete for funding from Project BioShield, a US programme that buys countermeasures against bioterror threats.

Other agencies — including the Environmental Protection Agency, the Food and Drug Administration, and the Centers for Disease Control and Prevention — get a total increase of \$100 million to help protect the nation's food and water supply against terrorist attacks, and to help fund the national stockpile of countermeasures.

Oceans aweigh

A welcome change may soon wash through ocean research. With its increased budget, the National Science Foundation aims to improve access to Arctic waters by building a new Alaska region research vessel. This would replace the ageing *RV Alpha Helix*. The vessel is expected to accommodate 500 researchers and students annually and be able to penetrate ice up to three feet thick. The agency is also planning an Ocean Observatories Initiative, to fund an interrelated network of stations, including those on the sea floor, on the coast and floating as buoys.

\$250 million on a controversial new programme to reprocess spent nuclear fuel and use it in reactors. Department officials say that the programme would reduce nuclear waste, but critics worry that the process is expensive and could lead to the proliferation of nuclear weapons (see *Nature* **439**, 509–510; 2006). "It is just incredibly misguided and ill-timed," says Paul Leventhal of the Nuclear Control Institute in Washington DC.

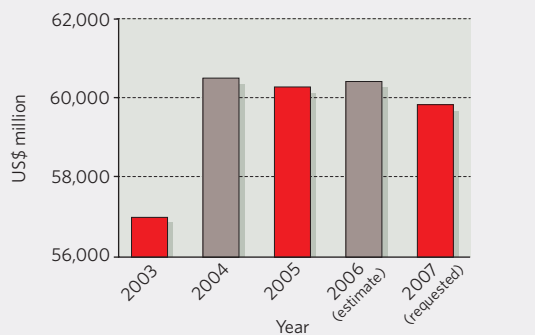
Less controversial are plans to increase funds for research on renewable energy. Solar enthusiasts are chuffed at a nearly 80% rise to \$148 million for work on photovoltaic cells, and fans of wind energy should see their research boosted by 13% to \$44 million. The request also provides increases for research on biofuels, hybrid batteries and hydrogen.

Meanwhile, the NSF would receive a 7.9% increase to \$6 billion under the president's plan. The boost would include major increases for information technology and infrastructure, and a contribution to the National Nanotechnology Initiative. Another new programme will fund grants for roughly 500 additional scientists, according to NSF director Arden Bement.

The budget request also calls for an increase of \$104 million in core funding to the National Institute of Standards and Technology. However, congressional earmarks eliminated by President Bush would cause the agency's overall budget to fall by 6% to \$535 million.

Physicists were overjoyed by the president's

FEDERAL SPENDING ON SCIENCE AND TECHNOLOGY



SOURCE: OFFICE OF MANAGEMENT AND BUDGET

newfound support. "This is great," says William Happer, who ran the Office of Science under the first president Bush. "The physical sciences have really been neglected for a long time." Mike Lubell, head of public affairs at the American Physical Society, adds that the request builds on a growing congressional push to increase funding for the hard sciences (see *Nature* **439**, 517; 2006). "Everybody seems to be seeing this the same way," he says.

The final decision on how much to spend now rests with Congress, and it remains to be seen whether budget-makers will see eye-to-eye. "The budget as a whole is not generous," says David Goldston, Republican chief of staff for the House Committee on Science. "But I would say that science programmes enter the fray in a strong position."

Geoff Brumfiel and Emma Marris, with additional reporting by Jacqueline Ruttiman
See pages 633 and 646 for more on the budget.

THE SCIENTIFIC BALANCE OF POWER

Show us the money

The world's most advanced economies are losing their scientific edge, some analysts claim. Fearing that weak research budgets will lead to weak economies, lawmakers on both sides of the Atlantic are preparing to pour billions of dollars into research and development (R&D; see pages 633 and 644).

But does the current financial reality reflect the doom-and-gloom predictions? To find out, *Nature's* reporters compiled this map of science and technology budgets in 19 key nations. We measured the total government investment in research, which includes both civil and defence R&D, and, to give a better sense of scale, we also looked at that investment as a proportion of each country's gross domestic product (GDP). These numbers do not take into account regional differences in the costs of doing research or any investment by private industry, but they do give a snapshot of how global leaders view priorities in science and technology.

The results confirm that, at least for now, the United States, Japan and western Europe have a commanding lead over the rest of the world. The United States, in particular, dominates global sciences, spending an astounding US\$134 billion on R&D this year alone — more than any other country or region. Much of that money, about 60%, is spent on defence-related R&D, but even when those funds are subtracted, the United States is still the standard by which all other nations measure themselves.

Things may not stay that way, however. Science budgets in the United States, Germany and France, have been stagnant in recent years, whereas budgets in Asian countries such as India and China have been racing upwards. If current trends continue, China is predicted to catch up with European Union spending, at least in terms of GDP share, by 2010.

Geoff Brumfiel

UNITED STATES

Getting physical

The US federal budget for R&D grew during the past decade, but biomedical research was the main beneficiary. Scientists in other disciplines have had to become resourceful in pursuing funding sources — a private donation was used to keep the nation's largest nuclear-physics experiment running this year, for example.

Current funding shortfalls are a threat to US competitiveness, warns a panel of experts at the National Academies. But neglected physicists may be in for a treat. In his 2006 State of the Union address, President Bush promised to double funding for the physical sciences over the next ten years.

US
133,780 \$
1.1 %

CANADA
6,320* \$
0.6 %

SOUTH AMERICA

Lagging behind

In the populist political climate of Latin America, science is often seen as more of a privilege than a priority. In Brazil, for example, a major block of funding for research was frozen last year to satisfy foreign creditors. And after several years of strong recovery in Argentina's R&D budget, funding is set to fall slightly this year, in favour of other priorities, including the national debt and infrastructure projects.

Nevertheless, specific initiatives are moving forward in both countries. Last April, Argentina unveiled a five-year, US\$10-million nanotechnology fund, and in September, Brazil revealed plans for a US\$5-million stem-cell research initiative.

BRAZIL
1,270* \$
0.2 %

ARGENTINA
410 \$
0.2 %

EUROPEAN UNION

An ambitious target

Taken together, the 25 countries that make up the European Union (EU) spend more than US\$80 billion per year on research, but they would like to spend more. For 2003, the EU set a formal goal to boost both government and industrial research spending from 2% to 3% of its collective GDP by 2010 (see page 640).

That would require not only billions in fresh government investment, but new funds from European industry. Given a sluggish European economy, industry is unlikely to defy past expectations and begin investing heavily in R&D any time soon.

The EU figures hide the fact that the individual nations move on different tracks. Some, such as Finland, already exceed the 3% target when spending by industry is included; others are taking steps to catch up with the EU average of just under 2%. The UK government, for example, boosted research spending by 13% last year. And Germany, Europe's largest economy, recently announced an additional US\$11 billion of tax money for science over the next four years. The newer member states, meanwhile, are doing their best to catch up from historic low levels.

ASIA

Rising stars

The collective research budgets of South Korea, China and India are less than one-quarter that of the United States, but their growth since the late 1990s has been rapid.

According to the latest figures available, South Korea's budget grew by 10% last year, China's budget expanded by 16% in 2004, and India's coffers swelled by an astounding 25% this year. The nations are also raising their profile with large projects such as the international fusion project ITER. And India is planning to create its own version of the US National Science Foundation with an annual budget of some US\$250 million.

But not all is rosy in the east. Japan's R&D budget has remained stagnant this year, largely because of rising healthcare and pension costs.



AFRICA

Slow progress

Of all the continents, Africa is perhaps farthest from realizing its full scientific potential. Progress has been hampered by regional conflicts, rampant inflation and weak governments. The instability led many African researchers to leave the continent in the 1990s, and many of those who remained have left science for other disciplines.

But there is some progress in South Africa, home to one of the region's most developed scientific communities. The government's R&D budget grew an impressive 24% between 2004 and 2005. And between 2004 and 2006, the government plans to spend some US\$23 million, about 10% of its annual budget, on technology programmes to alleviate poverty.

Government R&D budget in the current fiscal year (million US\$)
(* previous fiscal year)

Percentage of GDP

Source: OECD, UNESCO & individual governments

IN THE HANDS OF A MASTER

Fractal analysis has been used to assess the authenticity of paintings purporting to be the work of Jackson Pollock. **Alison Abbott** reports.

Jackson Pollock, famed for his 'poured' paintings, was defiant in facing down the cynics who viewed them as random splatterings. "I can control the flow of paint; there is no accident."

And several decades after the abstract expressionist's death, science proved him right. In the late 1990s, physicist Richard Taylor analysed a selection of Pollock's poured paintings and found they were composed of distinct fractal patterns — made by dripping or pouring paint straight on to a canvas. Indeed, it seems that 'Jack the Dripper' was refining the fractal characteristics of his paintings long before the mathematics to analyse them was invented.

Now, Taylor's evidence may prove critical in determining the authenticity of a group of recently discovered paintings that could be Pollocks. Given the financial implications, the physicist admits that he had to steel his nerves when writing his report. "This is a high-stakes game," says Taylor. "A Pollock poured painting can be sold for millions of dollars." In 1998, for instance, *Blue Poles: Number 11, 1952* was valued at US\$40 million. So Taylor knew that a negative result — if added to the doubts of Pollock experts — could strip many zeros from the value of the haul.

When the discovery of 32 possible Pollocks was made public in May last year it caused an immediate sensation. Pollock, an alcoholic who had a chaotic lifestyle and eventually died in a car crash, bartered several of his works for groceries. So it is likely that some of his paintings remain to be discovered. But although many claims have been made, only a small number of major poured paintings were formally authenticated before 1995 — when the Pollock-Krasner Foundation, set up under the will of Pollock's widow, Lee Krasner, disbanded its authentication board.

The provenance of the 32 paintings seems convincing. Alex Matter, son of the photographer Herbert Matter and painter Mercedes Matter, who were close friends of Pollock, found the works among a jumble of his parents' belongings. Labels in his father's handwriting identified them as paintings done by Pollock in the 1940s that he had acquired as 'gift + purchase'.

Alex Matter showed the works to the art dealer Mark Borghi, who in turn contacted Ellen Landau, a Pollock expert who had served on the Pollock-Krasner authentication board. Landau is now involved in preparing an exhibition called *Pollock Matters 2006*, which is being organized by Borghi and Matter to celebrate the 50th anniversary of Pollock's death.

The new poured paintings will play a key role in the show, and Landau will outline their significance in Pollock's career.

But other art historians have disagreed — some angrily — with the idea that the pictures are Pollocks. The doubters include Francis O'Connor, co-author of the definitive Pollock catalogue and another member of the authentication board, while it existed.

Given the high-profile dispute, and the large numbers of paintings involved — Taylor says they could represent up to a tenth of the significant poured paintings Pollock is known to have produced — the Pollock-Krasner Foundation decided it needed to get involved. It also decided that it required a more objective approach to authentication than the conflicting opinions of art historians could provide, especially if its judgement came to be challenged in the courts.

Traditionally, the authentication of a painting relies heavily on experts' visual assessments, supported by analysis of materials used in the work and knowledge of where it came from. "Art experts find it very stressful to make

"Pollock was honing his fractals a quarter of a century before they were defined."

judgements based on visualization alone," says Taylor. "They feel a bit let off the hook when materials or provenance can help."

But analysis of materials is of limited help in identifying true Pollocks, as the painter used common, off-the-shelf paints. And after *Life* magazine published a feature on Pollock and his sensational new approach to art in 1949, many readers tried their hand at his abstract expressionist style.

Although the provenance of the 32 paintings looks compelling, sceptics argue that the works could have been painted by Mercedes Matter, imitating Pollock's style. The fact that the poured paintings are on the type of board that Matter typically used, rather than Pollock's usual canvas, supports this argument, they say. The counter position is that Pollock

probably tried his method on Matter's boards because the two were so close.

So the foundation discreetly approached Taylor to act as a more objective arbiter, sending him six of the paintings to analyse. "From my point of view it was a good opportunity, as I was able to apply my research in the field," says Taylor.

Back in the late 1990s, Taylor, who has a degree in art theory as well as physics, decided to pursue his suspicion that Pollock's pouring technique could be described using fractal geometry. Fractal patterns, which repeat themselves at different magnifications, are often associated with chaotic systems. During the 1970s, mathematicians used chaos theory to reveal fractal patterns in natural objects such as coastlines, trees and flames.

There were two reasons to suspect that Pollock's paintings might obey fractal geometry. Moving around a large canvas laid on the ground, the artist let paint fly from all angles, using his whole body. Human motion is known to display fractal properties when people restore their balance, says Taylor, and films of Pollock seem to show him painting in a state of 'controlled off-balance'. Second, the dripping and pouring itself could be a chaotic process.

While continuing his research on nano-electronic devices (which display fractal patterns in their electrical properties), Taylor set about looking for fractals in five Pollock poured paintings in his spare time. He placed computer-generated grids over photographs of the works, and found two distinct sets of fractal patterns. One was on a scale larger than 5 cm; the other showed up on scales between 1 mm and 5 cm (R. P. Taylor, A. P. Micolich and D. Jonas *Nature* **399**, 422; 1999).

"Pollock was in control," says Taylor. The large-scale fractals are a fingerprint of the artist's body motion, he notes. "But the small-scale fractals are also to do with his choices — his height over the canvas, the fluidity of his paint, angle and force behind the trajectory, and so on."

Taylor also found that the fractal dimension of Pollock's works — a value that describes the complexity of a fractal pattern — increased through the years as the artist refined his technique. It seems that Pollock was honing his ability to generate fractals a full quarter century before fractal geometry was formally described.

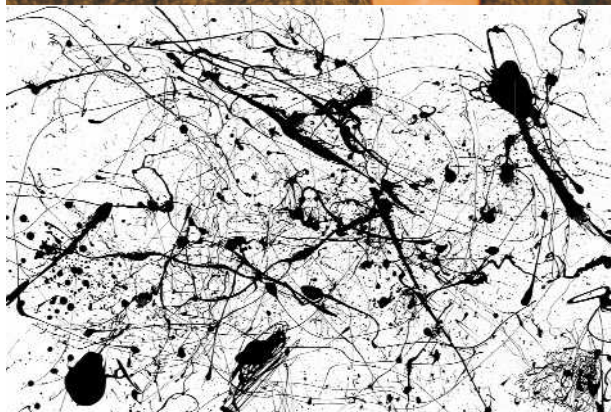
When he moved to the University of Oregon in Eugene in 2000, Taylor began a more comprehensive analysis to determine whether these fractal patterns were unique to Pollock. He used every last bit of information about the artist he could find, studying movements in a 1950 film of Pollock at work, and even the

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS



COURTESY OF J. FURTCHEIMER AND R. P. TAYLOR

PATTERN RECOGNITION LETTERS



Truth and beauty:
Richard Taylor compares the fractal characteristics of possible Pollocks with the real thing (see previous page) and student daubs (left).

splatters of paint that had missed the targeted canvas and landed on the floor, as recorded in old photographs. In total, he analysed 14 Pollock paintings, 37 imitations created by students at the University of Oregon and 46 poured paintings of unknown origin.

The genuine Pollocks had been painted in several ways — sometimes the artist flicked paint from a brush or stick; on other occasions he let the paint run down a brush or stick to fall on the canvas; sometimes he poured directly from the paint tin, or punctured holes in tubes of paints and squeezed them directly on to the canvas. Yet all had the same group of fractal characteristics. “The only shared thing in Pollock’s very different poured paintings is a fractal composition that was systematic through the years,” says Taylor.

Although the other poured paintings did include fractal patterns, none of them shared this particular group of fractal characteristics — and neither did Pollock’s accidental floor splatterings. Dan Rockmore, a mathematician

from Dartmouth College in Hanover, New Hampshire, has also searched for statistical signatures in art, in the works of painters such as Pieter Bruegel the Elder. He describes Taylor’s findings as “extraordinarily clever”.

The six paintings from the foundation arrived a year after Taylor submitted a paper describing these results to *Pattern Recognition Letters*, which has since been accepted for publication. Applying the same statistical techniques to these works, he found that none of them obeyed the fractal geometry he had observed in Pollock’s work. “I found significant deviations from Pollock’s characteristics,” says Taylor.

“Taken in isolation, these results are not intended to be a technique for attributing a poured painting to Jackson Pollock,” he wrote in a report to the Pollock–Krasner Foundation last July. “However, the results may be useful when coupled with other important information

such as provenance, connoisseurship and materials analysis.”

There were months of silence after Taylor delivered his confidential report. Given the high stakes for the art world, the foundation wanted to continue research on the provenance of each of the 32 paintings before giving a definitive thumbs-up or thumbs-down. Inevitably, rumours began to circulate — in the past few weeks, they have prompted Borghi himself to request Taylor to analyse a selection of the paintings. Borghi says he sees a lot of merit in what Taylor does, although he doesn’t believe authentication should rely on fractal analysis alone as painters often paint things out of style.

The foundation has now decided to go public with the results of Taylor’s study, while withholding a final, formal judgement. O’Connor, of the authentication board, pronounced his satisfaction that the findings “reinforce my own scepticism and reservations”. The foundation is choosing to pursue a consensus among Pollock experts; a draft statement seen by *Nature* calls Taylor’s results “a valuable contribution to our investigation”.

The results may be enough to cast doubt on the value of Matter’s

finds, at least until there is a final ruling from the foundation. Confidence in pattern analysis in art authentication is on the increase — which is partly why the foundation commissioned Taylor in the first place. And in the world of finance, whether it’s coffee, gold, or artworks, it is confidence that drives market prices. ■

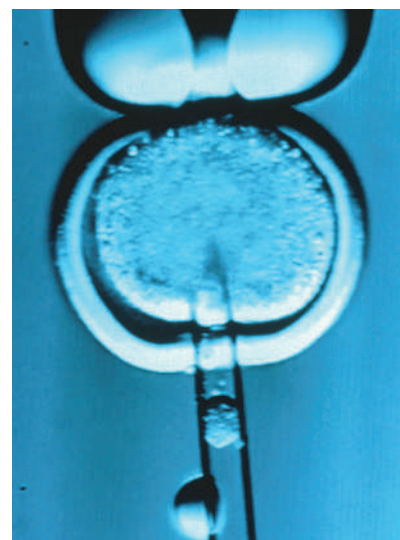
Alison Abbott is *Nature*’s senior European correspondent.

“I found significant deviations from Pollock’s fractal signature in the works.”

Cloning: what now?

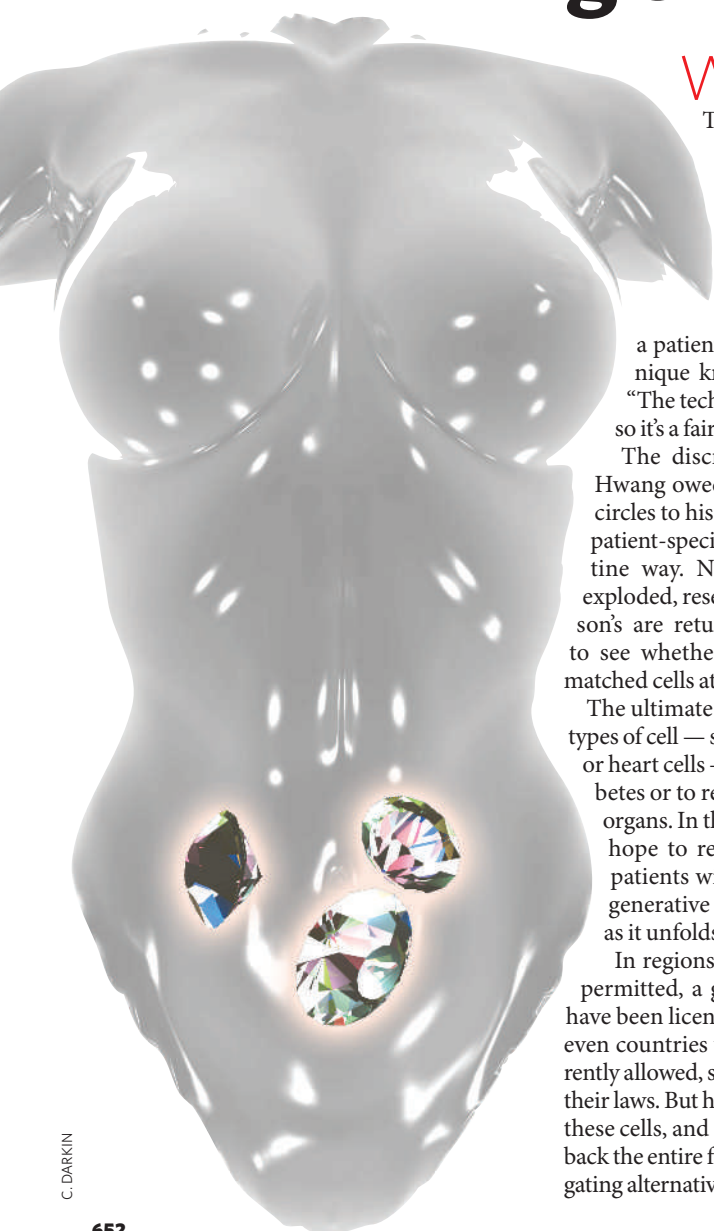
Less than a month ago, investigators at Seoul National University in South Korea announced that cloning researcher Woo Suk Hwang had lied when he claimed his team cloned human embryos with relative ease and produced stem cells from them.

The news was a significant setback for cloning researchers. In this special section, *Nature* looks at how biologists are regrouping. Below, **Carina Dennis** asks how they can get cloning to work given a very limited supply of eggs. **Phyllida Brown** looks at whether we will need therapeutic cloning at all, if immunologists can stop our bodies fighting transplants (see page 655). And on page 658, one of Hwang's closest rivals admits it may not continue its cloning quest.



J. KING-HOLMES/SPL

Mining the secrets of the egg



C. DARKIN

Women occasionally offer Alan Trounson their eggs. They approach the stem-cell researcher from Monash University in Melbourne, Australia, after he gives talks to patient focus groups. Trounson wants to treat neurodegenerative disease by using eggs to create cells that match

a patient's genetic make-up — a technique known as therapeutic cloning. “The technique is not legal in Australia, so it's a fairly brief conversation,” he says.

The discredited researcher Woo Suk Hwang owed his preeminence in cloning circles to his claims to have produced such patient-specific stem cells in an almost routine way. Now those claims have been exploded, researchers with aims like Trounson's are returning to the drawing board to see whether anyone can make patient-matched cells at all.

The ultimate dream is to create specialized types of cell — such as insulin-producing cells or heart cells — to treat diseases such as diabetes or to repair damaged hearts or other organs. In the nearer future, scientists also hope to recreate embryonic cells from patients with diseases such as neurodegenerative conditions, to study an illness as it unfolds and to test new drugs.

In regions where therapeutic cloning is permitted, a growing number of scientists have been licensed to start experiments. And even countries where the method is not currently allowed, such as Australia, are reviewing their laws. But human eggs are needed to make these cells, and a shortage of them could hold back the entire field. So researchers are investigating alternatives such as nurturing immature

eggs, growing artificial eggs in the lab and using animal egg substitutes. Each strategy comes laden with its own technical — and ethical — challenges.

To make therapeutic tissues such as heart cells, many researchers start with unspecialized, immature cells called embryonic stem cells. As their name suggests, such cells come from young human embryos, termed blastocysts, that are only a few days old. At the moment, researchers work on stem cells taken from surplus embryos created by clinics doing *in vitro* fertilization (IVF). But if these cells were simply transplanted into patients, the immune system would recognize them as foreign tissue and reject them (see ‘Do we even need eggs?’ on page 655).

Therapeutic cloning could, in theory, solve this problem. Working in animals, researchers have shown that if they transplant a nucleus from an adult body cell into an egg that has had its nucleus removed, the egg somehow ‘reprogrammes’ the adult nucleus back to an immature state, where it directs the development of an embryo. The resulting embryo is a genetic clone of the adult from whom the nucleus was taken. If this procedure works in humans, researchers could use cloned embryos to produce therapeutic or research cells that are essentially identical genetic copies of a patient's cells. Such cells should not be attacked by the patient's immune system.

In excess

But cloning is a wildly inefficient process, often requiring hundreds of eggs to produce a single viable clone. Indeed, one shocking revelation of the Hwang affair, was the sheer number of eggs his lab had got through¹. And obtaining human eggs is not easy. Donation is an unpleasant, invasive process that carries a small risk to a

woman's fertility and can, in rare cases, cause life-threatening side effects. This may make it hard to recruit donors. "We'll have to wait and see how difficult human eggs are to acquire," says Arnold Kriegstein, director of the Institute of Stem Cell and Tissue Biology at the University of California, San Francisco.

Most eggs currently donated to research are leftovers from IVF treatments — the ones that fail to fertilize and would otherwise be discarded. But these eggs typically fail to reprogramme², "probably for the same reasons they failed to fertilize," says Alison Murdoch, of the University of Newcastle Upon Tyne, UK. Murdoch and her team have successfully cloned a single blastocyst using excess eggs from women having infertility treatment³.

Ideally, researchers want healthy, competent eggs. Murdoch now asks women undergoing IVF treatment who produce plenty of eggs — more than 12 in a treatment — whether they would be willing to donate two eggs after the first dozen. "We have calculated that this does not significantly reduce their chances of a pregnancy," says Murdoch.

Greater good

Some researchers expect altruistic donations will be sufficient for research purposes. "My view is that most eggs are likely to come from women who have family members with a disease and want to donate their eggs to advance research on that disease," says Trounson.

But obtaining eggs for clinical use is likely to be a major obstacle, at least in the foreseeable future. "I can't conceive there will be enough eggs to use on a wide scale. In the end, we have no choice but to develop other methods," says Robert Lanza of Advanced Cell Technology. His company, based in Worcester, Massachusetts, has done therapeutic-cloning research using eggs from altruistic donors⁴ (see page 658).

And egg donations — especially those given altruistically — create an ethical quagmire. Is it appropriate to put healthy fertile women through such a procedure? Should they be paid for their eggs? These issues divide scientists. "Until we can get the efficiency to a reasonable level, we shouldn't be working in human eggs," says Stephen Minger, a stem-cell researcher at the Wolfson Centre for Age-Related Diseases in London, UK.

The most obvious place to start looking for alternatives to conventional donation is to go direct to the ovary. Although women only ovulate around 500 eggs in a lifetime, their ovaries are packed with thousands of eggs

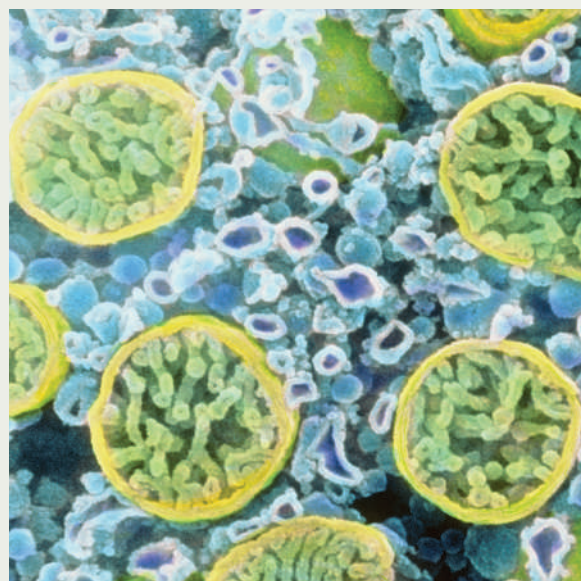
SPANNERS IN THE WORKS

Cloning is notoriously inefficient — it can take hundreds of eggs to produce a single embryo. Researchers trying to work out why can see one possible culprit: the tiny bodies that generate a cell's energy.

These powerhouses, called mitochondria and pictured right, have their own small genomes. But they also need to interact with the products of genes in the nucleus. A mismatch between mitochondrion and nucleus, or even between resident mitochondria and those introduced during the cloning process, could make the cell fail.

"Mitochondria are being totally overlooked — with devastating consequences," says Doug Wallace of the University of California, Irvine.

Fusing cells from two individuals might cause problems, says Wallace, but the biggest problems will arise if researchers try using animal eggs to reprogramme human



nuclei. "From our experience, combining the mitochondrial DNA from even a species as closely related as chimpanzees result in incompatibilities."

A recent study suggests that cellular function may not be

affected in cells derived from mice clones. But Wallace argues that such cells might be compromised when they move from the carefully controlled lab environment to repairing dilapidated organs inside a body.

"Human eggs are so precious — why practise on them?"

— Huizhen Sheng

at varying stages of development. What if researchers could somehow get hold of these — from ovarian biopsies, say — and grow them to maturity in the lab?

Biologists are making some headway in culturing eggs that are in the last stages of development. Outi Hovatta from the Karolinska Institute in Stockholm, Sweden, is working on eggs that are on the verge of being ovulated. Researchers are able to coax such eggs,

which are collected alongside mature eggs in normal IVF procedures, through the final stages of readiness for fertilization. Hovatta is optimistic that they could work in cloning experiments, which she is about to start. She estimates that altruistic donations of these almost-mature eggs would yield about 300 a year from a collaborating IVF clinic.

But culturing really immature eggs has proved extremely difficult. Egg development in humans is long, extraordinarily complex and not well understood. It begins in the embryo when special embryonic cells make their way to the developing ovary (see graphic, overleaf). Here, they divide many times to

generate millions of egg precursors, called primary oocytes. At birth, the ovary contains about half a million follicles: these consist of a primary oocyte wrapped in one or more layers of cells that support the oocyte as it grows and accumulates nutrients needed for the early development of an embryo.

Only after puberty do follicles fully develop, with one follicle growing to full size per menstrual cycle and releasing its enclosed oocyte. Just before ovulation, this oocyte ejects half its chromosomes, getting rid of half the remainder when a sperm makes contact with its own genetic cargo.

In the hands of John Eppig, a researcher at the Jackson Laboratory in Bar Harbor, Maine, culturing eggs through these stages looks almost easy. He can create live mice pups by fertilizing eggs that have been cultured in the lab from ovaries extracted from newborn mice⁵. Although the first offspring of these experiments — dubbed Eggbert — was a sickly creature, subsequent mice look healthy.

But it is a different story with larger animals. "It's tougher in species other than rodents because it takes so much longer for egg development to occur," says Eppig; it takes

more than three months for a human egg to mature. Another factor is the relatively gargantuan size of a human egg, which swells to more than 100 micrometres. The challenge is to ensure that the voluminous egg receives adequate nourishment, as well as figuring out the right factors to coax it along the development pathway.

That said, researchers have had some success with human eggs. Ronit Abir from the Rabin Medical Center at Beilinson Hospital, Israel, has nurtured isolated follicles for several weeks *in vitro*. Controversially, she has also cultured immature eggs from aborted human fetuses to almost the same stage⁶. Aborted fetuses are not likely to be a good source of eggs, given the obvious ethical concerns, including the fact that a fetus cannot give its consent. But Abir says the work could unravel the mysteries of culturing eggs from early development and reveal ways to restore the fertility of cancer patients who have had their ovaries extracted and frozen.

Hovatta's team has been able to grow human primary oocytes in intact ovary slices in culture and has nudged them along several developmental stages. But these efforts to culture primary oocytes have yet to yield eggs that can be fertilized. "In the beginning, I thought it would change everything, but now I see how slow the rate of progress is," says Abir.

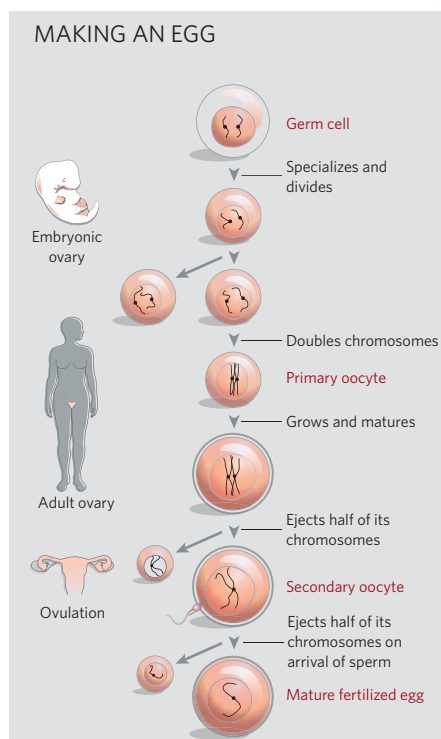
Make it up

Others are going right back to the earliest stages and trying to develop eggs from scratch using embryonic stem cells^{7,8}. If such cells are left to grow very densely on a culture dish under the right conditions, they clump together and, amazingly, will form egg-like structures.

Researchers have not yet shown that these egg-like cells can be fertilized. But they might be good enough to reprogramme a nucleus, according to Hans Schöler of the Max Planck Institute for Molecular Biomedicine in Münster, Germany. Schöler pioneered the growth of egg-like cells from the embryonic stem cells of mice. The research field is anxious to see whether reprogramming will be possible — Schöler claims to have made one unsuccessful attempt. "We are still working out the conditions," he says. "It's not as trivial as we'd thought."

With human eggs presenting so many difficulties, some researchers are exploring the possibilities of animal eggs, at least for research purposes. "Human eggs are so precious — why waste them to practise on?" asks Huizhen Sheng, from the Centre for Developmental Biology at Xinhua Hospital in Shanghai.

Sheng's lab sparked an international storm when the media reported in 2002 that she was



using rabbit eggs to clone human blastocysts. Some newspapers ran headlines about animal-human monsters, fuelling public hysteria. The debate was reignited recently when Chris Shaw, a neurologist at King's College London and Ian Wilmut, the creator of cloned sheep Dolly, based at the University of Edinburgh, announced that they were seeking approval to do similar experiments.

Sheng has published her data⁹, but the research community remains unconvinced that her method works. "You have to be uncertain of that work until it is repeated," says Trounson. Sheng attributes certain discrepancies to

the lab's culturing methods, a problem that she says has now been rectified.

Creating hybrids of human cells and animal eggs is banned in many countries, under review in others such as Australia, and yet to be tested in the more permissive regulatory environment of Britain. But with scant supplies of fresh human oocytes, many researchers see animal eggs as the only practical alternative for refining therapeutic-cloning techniques. And they could be useful for generating patient-specific lines to study the genetic basis of human diseases. "We are trying to understand disease processes to identify new therapeutic targets. These cells are not for putting back into people," says Shaw.

Mismatched

Still, many scientists are doubtful that animal eggs will yield useful human embryonic stem-cell lines. The main concern centres on mitochondria, the bacteria-like powerhouses of the cell. Mitochondria have their own genomes, which interact with the genome in the cell's nucleus. Mixing nuclei and mitochondria from different species simply may not work (see 'Spanners in the works', previous page). "It's hard enough to keep our nuclear chromosomes in sync with our own mitochondrial DNA and we're the same species," says Irving Weissman of the Stanford School of Medicine, California. And yet scientists are undeterred. "It's an important question and difficult to answer until you have done the experiments," says Shaw.



Slim pickings: it is hard to harvest quantities of human eggs because the process is unpleasantly invasive.

Given that eggs are so problematic, some teams are attacking the problem of reprogramming from a different angle. They are trying to see whether other kinds of cells share an egg's ability to reprogramme a nucleus. One such candidate is an embryonic stem cell itself. Recently, Harvard University researcher Kevin Eggan and his colleagues transformed adult body cells to an embryonic state by fusing them with embryonic stem cells¹⁰. And some experiments have even suggested that embryonic stem cells might be better at certain aspects of reprogramming than oocytes.

But the major drawback of this method is that the chromosomes of the embryonic stem cell used to spark the process are retained. This limits a cell's therapeutic potential because a patient's immune system could recognize the leftover chromosomes and launch an attack. Researchers are working on fixes, however. For example, Paul Verma from Monash University has devised a way of getting rid of the unwanted chromosomes¹¹, and now has unpublished evidence that mouse cells might be reprogrammed using this approach.

Others are searching for those seemingly magical factors in eggs that allow them to wind an adult nucleus back to an embryonic form. Nobuaki Kikyo of the Stem Cell Institute at the University of Minnesota in Minneapolis, for example, has fished out factors from frog eggs that can repack chromosomes, dismantle the nucleus's structure and switch on gene activity — all key aspects of reprogramming¹². But this approach will take time. "Someone might get lucky, but I think it's a long way off," says Keith Latham from Temple University School of Medicine in Philadelphia, Pennsylvania.

There is unlikely to be one single way to mimic the almost mystical reprogramming ability of a human egg. The answer, says Weissman, could be to combine methods — kick-starting the process with one approach, and finishing it with another. ■

Carina Dennis is Nature's Australasian correspondent.

1. Cyranoski, D. *Nature* **439**, 122–123 (2006).
2. Lavoie, M.-C., Weier, J., Conaghan, J. & Pedersen, R. A. *Reprod. Biomed. Online* **11**, 740–744 (2005).
3. Stojkovic, M. et al. *Reprod. Biomed. Online* **11**, 226–231 (2005).
4. Cibelli, J. B. et al. *J. Regen. Med.* **2**, 25–31 (2001).
5. O'Brien, M. J., Pendola, J. K. & Eppig, J. J. *Biol. Reprod.* **68**, 1682–1686 (2003).
6. Biron-Shental, T. et al. *Fertil. Steril.* **81**, 716–719 (2004).
7. Hübner, K. et al. *Science* **300**, 1251–1256 (2003).
8. Clark, A. T. et al. *Hum. Mol. Genet.* **13**, 727–739 (2004).
9. Chen, Y. et al. *Cell Res.* **13**, 251–263 (2003).
10. Cowan, C. A., Atienza, J., Melton, D. A. & Eggan, K. *Science* **309**, 1369–1373 (2005).
11. Pralong, D. et al. *Cloning Stem Cells* **7**, 265–271 (2005).
12. Tamada, H. et al. *Mol. Cell. Biol.* **26**, 1259–1271 (2006).



C. DARKIN

Do we even need eggs?

Nobody likes rejection, but for a transplant patient it can be a death sentence. The risk that a patient's immune system will see a transplanted organ, or graft, as 'foreign' rather than 'self', forces transplant patients on to immunosuppressant drugs that can have severe side effects. Therapeutic cloning, its enthusiasts say, could solve the problem by allowing doctors to grow cells and tissues that are perfectly matched to individual patients. In this approach, a patient's DNA is transferred into an egg which is persuaded to develop into stem cells that in turn generate spare-part tissues. But many researchers now think therapeutic cloning is unrealistic, largely owing to the scarcity of human eggs.

So the spotlight is turning on to different strategies, aimed at persuading the immune system to tolerate foreign tissue. "The field is moving very fast," says Harry Moore of the Centre for Stem Cell Biology at the University of Sheffield, UK. "Ten years ahead there may be no need for cloning, except in certain cases."

There are different ways to increase the success of tissue transplants. One is to develop generic stem cells, cell lines and tissues, and then persuade the immune system to accept them. These could be therapeutic transplants of, say, insulin-producing cells to treat diseases such as type 1 diabetes. Those championing this idea admit it is years, or decades, from the clinic. So more pragmatic

approaches are also under development. Instead of trying to make the immune system perfectly tolerant of a transplant, some researchers are aiming to increase its tolerance enough for patients to sharply reduce their dependence on powerful immunosuppressive drugs.

At the moment, most people who have an organ transplant face a lifetime of treatment with drugs that affect the whole immune system, such as cyclosporine and steroids. Although these drugs increase the life of a grafted organ by several years, they often fail to prevent its eventual rejection, and they put patients at risk of infections, cancers and kidney failure.

Drug problem

In studies on small numbers of patients who have had organ transplants, medical teams are discovering that they can manipulate the immune system so that drug treatment can be reduced. For example, Chris Watson and Roy Calne at Addenbrooke's Hospital, Cambridge¹, are among those who gave patients an antibody, Campath-1 (alemtuzumab), at the time of an organ transplant. This antibody depletes lymphocytes, a large family of key immune cells.

Watson and Calne then gave lower doses of immunosuppressive drugs to the patients than would normally be given following a transplant, and no steroids. For the five-year study period, the patients' grafts survived as

"Ten years ahead there may be no need for cloning."

— Harry Moore

well as did those in patients given conventional treatment. This treatment approach is not ready for widespread use, says Herman Waldmann of the University of Oxford, UK, who developed Campath-1. But, he adds, it demonstrates the principle that the use of immunosuppressive drugs can be minimized.

Although researchers disagree about exactly how the antibody treatment works, some speculate that by holding off an immune attack, it somehow gives the immune system time to 'learn' to tolerate the graft. The aim now is to understand more about how this tolerance develops, so that more precise therapies can be designed. After all, says Waldmann, an antibody that blocks whole classes of lymphocytes is still a relatively crude instrument.

The immune system distinguishes between 'self' and 'non self' proteins because of the actions of a subset of lymphocytes called T cells. Each T cell carries a receptor on its surface, uniquely shaped to fit a specific protein fragment or peptide. Other cells of the immune system patrolling the body pick up proteins, or antigens, and present fragments to the T cells. If a fragment fits into a specific T-cell receptor, it acts rather like a key in a lock, and, given certain conditions, will switch the T cell into an active state. The T cell then coordinates an immune response against that protein.

Of course, if the T cells became activated in response to self proteins, disaster would ensue. So when new T cells maturing in the thymus (an organ just above the heart) encounter a range of protein fragments,

T cells that specifically recognize self antigens are killed off or 'deleted'. This editing, known as central tolerance, is what usually protects us against autoimmune disease. The process happens mainly in infancy and youth, with the thymus shrinking as an individual reaches maturity.

But while this central tolerance is developing, the body can be fooled into accepting foreign cells as self — at least in newborn mice². The young animals are given infusions of bone-marrow cells. These cells are the precursors of T cells and dendritic cells, another type of lymphocyte that plays a key role in presenting protein fragments to T cells. When the dendritic cells derived from this foreign bone marrow migrate to the thymus and present their protein fragments to the T cells maturing there, the fragments are seen as self and any T cells that recognize them are killed off. As a result, the animals later accept grafts of skin from the same donor source.

Layers of command

But this approach is obviously not suitable for treatment in adult humans, because the thymus is producing so few new T cells. So researchers have tried to manipulate the mature immune system to achieve a sort of artificial central tolerance. In animal experiments, the host's bone marrow is partly inactivated and new donor bone marrow infused in

its place. The presence of the foreign bone-marrow cells creates a state known as 'mixed chimaerism'³, where T cells that react to either host- or donor-derived antigens are killed off, allowing the animal to accept other tissue from the same donor.

Megan Sykes, David Sachs and colleagues have tried to harness this phenomenon of chimaerism to persuade the immune system to

accept transplants of bone marrow and kidneys in a small number of adult patients. They treated the patients with drugs to deplete the host lymphocytes, then transplanted the new tissues. At first, the patients' blood appeared chi-

maeric — that is, the researchers found various types of blood cells in it bearing the donor's signature. If these findings reflect the animal experiments, they would suggest that T cells reactive to the donor tissues had been deleted as though they were targeting self.

But there was a puzzle. The grafts survived for years, yet, strangely, the donor blood cells faded away after about 12 weeks, suggesting the patients were no longer chimaeric. This suggested that chimaerism could not fully explain the survival of the grafts, but that something else must be helping to protect them from rejection. Sykes suggests that the kidney grafts themselves must be doing something that helps the host immune system tolerate them. Instead of involving the central deletion of T cells in the thymus, this form of

"The important thing is to take the field onwards in all the areas with potential."
— Herman Waldmann

GREAT EXPECTATIONS

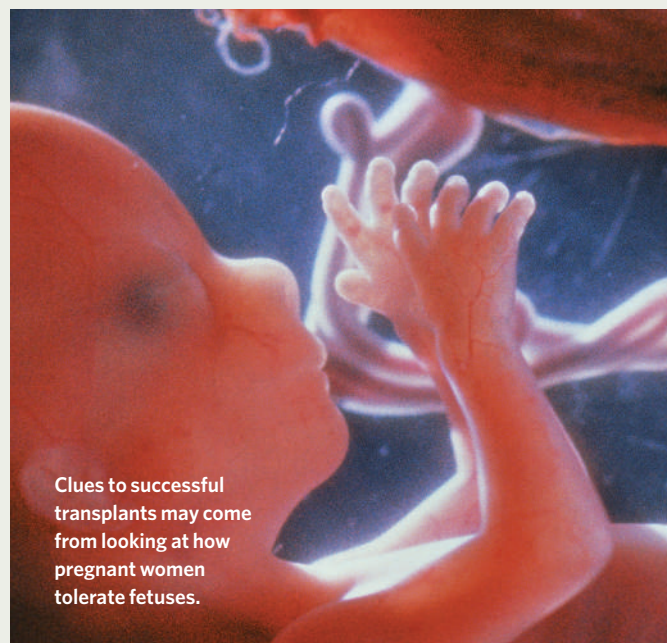
Millions of people around the world happily tolerate foreign tissues for months on end. No scientific trick here: simply pregnancy. Researchers are trying to understand how pregnant women tolerate their fetus in the hope of understanding tolerance, or even preventing miscarriage.

One such team is headed by Harry Moore at the University of Sheffield, UK. He and his team are studying a particular kind of blood-vessel cell found in the placenta. This secretes a protein known as HLA-G, which seems to help prevent the mother's T cells from attacking her fetus.

Recently, researchers found that a type of immune cell called a monocyte can also make HLA-G. Intriguingly, these cells infiltrate

newly transplanted tissue. The higher the levels of HLA-G expressed, the greater the chance that the transplant will be accepted. This finding, says Moore, suggests that HLA-G may naturally help to protect some tissue grafts, although he stresses that no one knows how the process works.

Moore's group hopes that studies of this kind of blood-vessel cell may reveal more about how an embryo normally implants and why some mothers' immune systems do reject their fetuses. In the longer term, says Moore, it might be possible to develop the cells for transplant therapy in heart disease. The cells might also be given with other therapeutic tissues, with the aim of making the recipient tolerate both types of cell.



Clues to successful transplants may come from looking at how pregnant women tolerate fetuses.

EDELMANN/SPL

tolerance seems to be happening in the periphery, around the grafted tissue.

One possibility, she suggests, is that a specialist set of T cells known as regulatory T cells is involved. Often found lurking in the vicinity of tolerated grafts, regulatory T cells are beginning to attract interest from many immunologists. Waldmann thinks that regulatory T cells may be key players in maintaining the long-term survival of a graft. He hopes that these cells could eventually be harnessed for preventive treatments that would dampen down the immune response to a graft, and has coined the concept 'negative vaccination'.

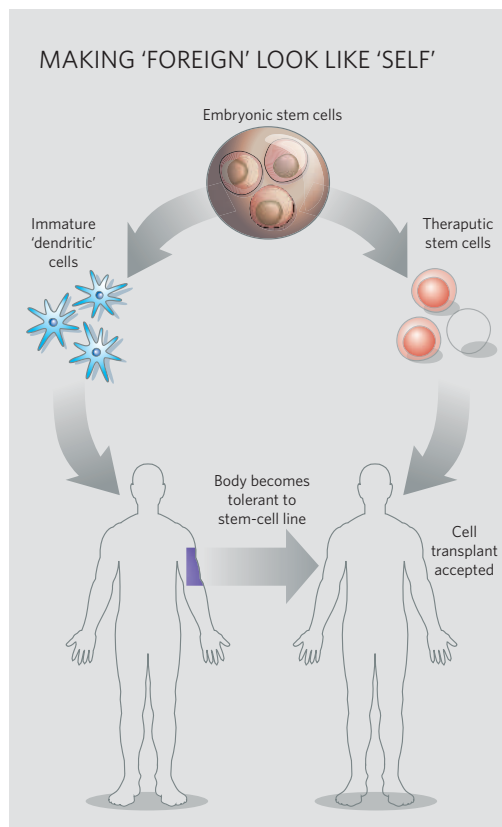
Clever cultures

Regulatory T cells seem to control and police the activity of other immune cells in their vicinity. Waldmann's idea that they protect grafts by suppressing aggressive immune responses to the donor antigen depends on the role of dendritic cells.

If T cells are the generals commanding the immune army, then dendritic cells are their sentinels and scouts. They patrol sites, such as the skin and mucous membranes, picking up foreign proteins and presenting fragments of them to T cells. The dendritic cells can also act as arbitrators on the fate of these proteins, influencing whether the T cells attack or tolerate them⁴.

If the dendritic cells encounter the foreign tissue in the presence of 'danger signals' — other molecules that trigger inflammation such as bacterial sugars or self proteins that are only released around wounds — they rapidly switch to a mature form. This form of dendritic cell produces signals that mobilize the T cells to attack. But in the absence of this signal, dendritic cells remain in an immature form. They still present the foreign protein to T cells, but will usually 'decommission' the T cells that specifically recognize the antigen, preventing them from responding aggressively to it.

Waldmann holds that the interaction between dendritic cells and T cells in the vicinity of the transplant may be crucial to allowing the graft to be tolerated. At the time of transplantation, the tissue is inflamed through local trauma, but lymphocyte-depleting drugs stop an acute attack on it by the host immune system. This may create a window of time, argues Waldmann, during which the tissue can heal. Once it has healed, and provided there are no infections, then dendritic cells presenting foreign antigens to T cells around the graft will do so without a danger signal. As a result, the dendritic cells will not activate the T cells they



form most likely to make T cells tolerant.

Fairchild reasoned that human dendritic cells, cultured in the same way, could play a role in inducing transplant tolerance. If the dendritic cells and the therapeutic stem cells required for a given transplant — say, heart-muscle cells — could both be cultured in lines from the same stem-cell source, the dendritic cells could be used to make the recipient tolerate the therapeutic cells⁶ (see graphic).

Early days

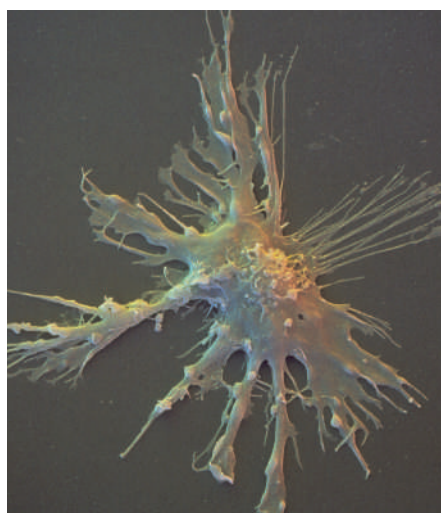
Fairchild and his colleagues have yet to test this idea fully in mice, although their early unpublished experiments suggest that mice injected with dendritic cells carrying a specific protein can be made to accept skin grafts carrying the same protein. Fairchild has begun culturing human dendritic cells from embryonic stem cells.

There is, however, the risk that a dendritic cell could switch into its mature state, and so activate T cells. If this were the case, the cells could help destroy, rather than protect, the graft. Fairchild and Waldmann are exploring the idea of treating the cells with various compounds in culture to keep them in their 'tolerizing' form⁷. But, as Fairchild stresses repeatedly, these are early days.

But not all researchers are confident that cells differentiated from embryonic stem cells will end up being suitable for transplant therapy. Mick Bhatia, an immunology researcher now at McMaster University in Ontario, Canada, worries that even under the most rigorous conditions, an undifferentiated stem cell could still slip in among the differentiated cells. Undifferentiated stem cells can trigger a form of cancerous growth. Waldmann and Fairchild agree that before the approach could be used for treatment, tests must be developed to identify any undifferentiated cells in a cultured line.

No one pretends that tolerance is going to arrive tomorrow. "But the important thing is to take the field onwards in all the areas with potential," says Waldmann. As degenerative diseases such as heart disease become more common, Fairchild and other researchers are convinced that they should intensify their efforts. "We've got a lot further to go, but it's very exciting."

Phyllida Brown is a science writer in Exeter, UK.



Dendritic cells could be harnessed as a tool to help promote tolerance of tissue transplants.

1. Watson, C. J. et al. *Am. J. Transplant.* **5**, 1347–1353 (2005).
2. Billingham, R. et al. *Nature* **172**, 603–606 (1953).
3. Buhler, L. H. et al. *Transplantation* **74**, 1405–1409 (2002).
4. Steinman, R. M. et al. *Annu. Rev. Immunol.* **21**, 685 (2003).
5. Fairchild, P. J. et al. *Curr. Biol.* **10**, 1515–1518 (2000).
6. Fairchild, P. J. et al. *Trends Immunol.* **25**, 465–470 (2004).
7. Fairchild, P. J. et al. *Intl. Immunopharmacol.* **5**, 13–21 (2005).

BUSINESS

No end in sight for stem-cell odyssey

Advanced Cell Technology plans to tread cautiously on the ground left vacant by the collapse of a South Korean scientist's claims to have cloned a human embryo, as David Cyranoski reports.

Little more than two years ago, one of the research teams considered most likely to clone a human embryo was at Advanced Cell Technology (ACT), a small biotechnology company in Worcester, Massachusetts.

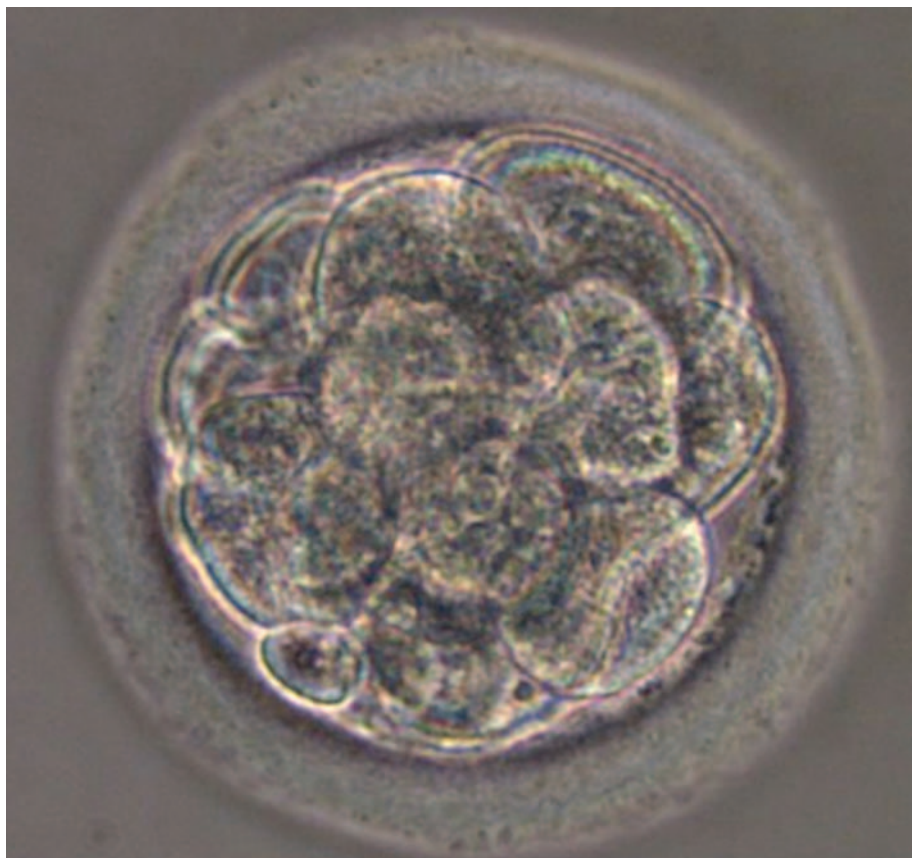
ACT — a company whose rocky history had already carried it from work on cloning farm animals into human embryonic stem-cell research — had created cloned early-stage human embryos, called morulae. In the view of Robert Lanza, its vice-president for research, it was well placed to take that work to the next stage, and clone an embryo. “We were so close,” he says. “It would be nice to finish up.”

But towards the end of 2003, the company learned that Seoul National University's Woo Suk Hwang, then an animal-cloning specialist little known outside South Korea, was about to announce that he had succeeded both in cloning a human embryo and in deriving a line of stem cells from it. Hwang's results (W. S. Hwang *et al. Science* **303**, 1669–1674; 2004) were later found to be fabricated, and were retracted last month. Nonetheless, their appearance was a blow from which ACT's research plans may never recover.

ACT was founded in 1994 and initially specialized in the reproductive biology of farm animals. But the company took a sharp turn into human reproductive biology in 1999, when Michael West, founder of the biotechnology company Geron, joined the company as chief executive, with Lanza, a specialist in regenerative medicine, as his research chief. ACT was soon involved in a variety of projects to develop human stem cells.

But only stem cells produced from an embryonic clone of a patient's cell would be an identical match to the individual's genetic make-up and safe from the possibility of rejection (see News Features, pages 652 and 655). So despite the controversy that surrounded the field in the United States, ACT decided to apply somatic cell nuclear transfer — the technique sometimes known as therapeutic cloning — to humans. “We knew it was extremely controversial, but considering the medical potential, we thought it was appropriate,” recalls Lanza. “The controversy turned out to be greater than we expected,” he ruefully admits.

By 2001, ACT had reached what it saw as an early landmark on the road to generating



ACT

Nearly there? A cloned human morula produced by somatic cell nuclear transfer at ACT in 2003.

cloned human cells. It had managed to clone very early-stage embryos, containing six human cells (J. B. Cibelli *et al. J. Regen. Med.* **2**, 25–31; 2001). But the paper didn't impress developmental biologists. “Many scientists thought it was a trivial publication of essentially failed experiments,” says one cloning specialist, who declined to be identified. Lanza disputes that characterization. “They were early data,” he admits. “But rumours were getting around, and we needed to publish.”

On the brink

By October 2003, Lanza says, greater progress had been made. Thirteen of sixteen cloned human eggs activated by transfer of a somatic nucleus had developed to morulae, a more mature stage from which, however, researchers were unable to derive stem-cell lines. In follow-up experiments, he adds, they

identified problems with the medium used to culture the cells, which might have explained the failure to cultivate stem-cell lines. Lanza claims that they were on the brink of getting the experiment to work.

But when word came out of the South Korean group's publishing plans, Lanza says, the work was shelved. “We were seen as the losers,” he recalls bitterly. As well as cash, ACT's already limited supply of donated human eggs dried up. “We couldn't afford pipettes,” says Lanza, “much less eggs that cost thousands of dollars.”

Lanza became one of the few scientists to publicly challenge Hwang's research. In an interview with United Press International published on 12 February 2004, he demanded that the Korean results be independently verified by means of forensic DNA tests. But his complaint — coming as it did from one of

Hwang's fiercest rivals — fell on deaf ears.

Now that Hwang's work has been retracted, however, ACT is ill-positioned to win back the initiative. According to William Caldwell, the seasoned corporate rescue specialist who became ACT's chief executive in January 2005, the company's top priority is instead to find sources of revenue. "With what modest funds there are, the company should develop products rather than do basic science," he says.

In January 2005, the company raised \$8 million in venture capital, and immediately afterwards, it undertook a 'reverse buyout' — a financial manoeuvre whereby a listed company that has scant operations (in this case, Two Moons Kachinas, a Utah-based maker of native American dolls) takes over an unlisted one (ACT) in order to obtain listed status for the latter. For a company that needs to raise money, that's an easy way to get onto the stock market without a time-consuming and costly initial public offering. The company raised another \$18 million in September. Today, says Caldwell, ACT is in "its most liquid state ever".

Investors impressed

Caldwell says that investors are impressed by ACT's dedication to medical applications. One technology aims to coax human embryonic cells to become retinal pigment epithelium that could be used to treat macular degeneration (I. Klimanskaya *et al.* *Cloning Stem Cells* **6**, 217–245; 2004). ACT is also investigating how to get stem-cell treatments that are not matched to a patient's DNA accepted by the body's immune system.

Lanza remains hopeful about the long-term potential of human cloning. His team has recently shown that cloning in mice does not necessarily mean destroying the embryo (I. L. Weissman *Nature* **439**, 145–147; 2006). He claims that the technique is so efficient that he might need only five or ten eggs to clone a human embryo — against more than 2,000 used by Hwang's team.

But his optimism about human cloning by somatic cell nuclear transfer is not widely shared. "I'm not aware of any companies besides ACT doing it," says venture-capitalist Douglas Fambrough of Oxford Bioscience Partners in Boston, Massachusetts. A leading US stem-cell scientist, who didn't want to be identified, was more scathing. "I am doubtful that patents on human somatic cell nuclear transfer will have any significant commercial value," he says. "That's why I roll my eyes every time I hear Lanza claim that Hwang's work derailed their company's fund-raising."

And Caldwell certainly won't be betting ACT's future on such experiments bearing fruit. "We are assessing what others are doing," says Caldwell, "but we won't throw major resources at it."

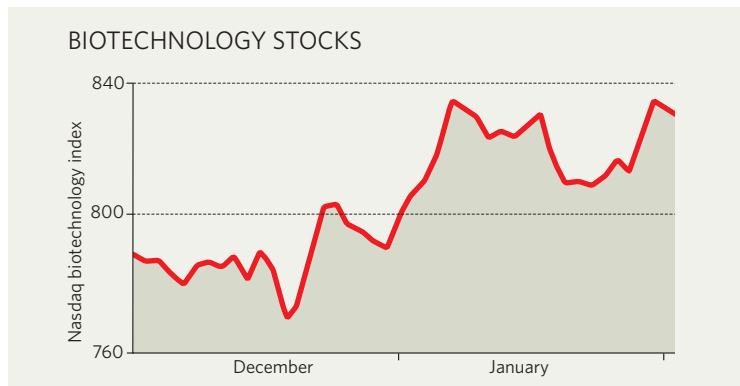
IN BRIEF

ROCHE REBOUND Swiss drug giant Roche reported sales surging by 20% in 2005, to reach 35.5 billion Swiss francs (US\$27.8 billion). The numbers were driven by the Basel-based company's portfolio of cancer drugs, whose sales rose by 42% — and by strong demand for Tamiflu, the influenza treatment that countries are stockpiling in preparation for a possible pandemic. Sales of the drug almost quadrupled, to 1.56 billion Swiss francs.

MIGHTY WIND Power companies in the United States erected enough wind turbines in 2005 to supply 2,500 megawatts of electricity, says the American Wind Energy Association (AWEA), based in Washington DC. The trade group predicts the installation of another 3,000 megawatts of wind power in 2006 — a big increase on the previous record level of installation, which was 1,700 megawatts in 2001, but still less than half the capacity installed in Europe last year. The AWEA wants Congress to extend a subsidy arrangement that encourages utilities to install wind power.

EYES ON THE PRIZE A Boston biotechnology company devoted to discovering novel combinations of existing drugs announced a collaboration aimed at exploiting the market for eye-disease medications. CombinatoRx (see *Nature* **439**, 390–391; 2006) has signed up Fovea Pharmaceuticals of Paris, France, a new, privately held biotechnology firm, to fund and conduct preclinical, early and mid-stage clinical development of combinations of eye drugs from the Boston company's proprietary database. Fovea will receive exclusive commercial rights to the resulting products in Europe.

MARKET WATCH



This week Wood Mackenzie, an Edinburgh-based research and consulting firm, reviews recent trends in biotechnology stocks.

After a flat close to 2005, the Nasdaq biotechnology index moved forward in January, with most of the rise attributable to a flurry of recent clinical trials results and acquisition activity.

Arena Pharmaceuticals of San Diego, California, for example, revealed promising interim phase II data for its obesity drug candidate APD356, and saw its shares rise by almost two-thirds. Massachusetts-based Vertex Pharmaceuticals also enjoyed a 42% rise, driven by strong clinical data from its hepatitis C drug candidate, VX-950.

Offsetting these gains were some spectacular falls: Rigel Pharmaceuticals of San Francisco, California, saw its share value nosedive by 64% after disappointing phase II data for its nasal allergy candidate R112. And

worse-than-expected efficacy data for anticancer drug candidate G-024856 saw shares in Massachusetts-based Curis lose a quarter of their value.

Despite reporting rapid revenue and profit growth in 2005, shares in industry bellwether Amgen, of Thousand Oaks, California, has slipped by 12% in the past eight weeks. Given Amgen's heavy weighting in this index, it is remarkable that the biotechnology index still managed to rise by 4% over the period. But Amgen itself played a part in that, by announcing plans to acquire one of its important partners in drug development, biotechnology company Abgenix of Freemont, California, for \$2.2 billion.

Shares in Abgenix, whose main product, the antibody Panitumumab, is expected to win US approval for the treatment of colorectal cancer early this year, rose by 67% on the news. ■

SOURCE: NASDAQ

Climate may not be linked with circulation slowdown

SIR — Your News Feature “A sea change” (*Nature* **439**, 256–260; 2006) states that evidence for the huge effects on climate of past thermohaline shutdowns is “near indisputable”. You then claim that the best such evidence is the coincidence of thermohaline slowdown with the flooding of the North Atlantic following the collapse of Lake Agassiz, about 12,000 years ago at the beginning of the Younger Dryas cold period.

Yet Wallace Broecker, one of the chief proponents of the relationship between thermohaline circulation changes and climate, finds otherwise. In a recent study (T. V. Lowell *et al.* *Eos* **86**, 365–373; 2005), Broecker and colleagues suggest that the case for the coincidence of these events is quite weak, and might well be wrong. Instead, they say, “preliminary findings imply a retreating ice sheet margin approximately 1000 years younger than previously thought, which would have blocked key meltwater corridors at the start of the Younger Dryas”.

Of course, the proximal cause of a thermohaline “shutdown” (if such a thing can happen) is a separate issue from the influence of such a shutdown on climate. But it is highly relevant to this discussion of the sensitivity of the thermohaline circulation to current and future climate forcing.

Eric Steig

Department of Earth and Space Sciences,
University of Washington, Box 351310, Seattle,
Washington 98195, USA

No unfairness in funding of Croatian minister's project

SIR — Your News story “Croatian scientists call for openness over funding” (*Nature* **439**, 7; 2006) brings attention to the funding of technology-development projects in Croatia. The petition discussed in this story was signed by only 26 scientists inside Croatia (the other signatories are foreign or students), and in my view it is frivolous.

The funding application made by Krešimir Pavelić of the Rudjer Bošković Institute, as senior investigator, and Dragan Primorac, who was then at Holy Spirit II Hospital, was among 28 applications received for this ‘nucleus’ programme. After an independent review including international experts, eight applications were recommended for funding. The technology council that reviewed the Pavelić–Primorac application was appointed by Gvozden Flego, the minister in the previous government. Primorac was the most junior of the 13 members of the council; as is customary and appropriate, he excused himself when his application was considered.

The recommendation for funding was forwarded to Flego and was approved before Primorac became a member of the new government. Primorac did not affect the decision about the funding of his own project. He stepped down from the project when Pavelić assumed the full directorship of it, and before any funds were released to the Rudjer Bošković Institute.

The technology council closely monitors the execution of projects and requires quarterly reports from project directors. This mechanism assures that the funds are spent in accordance with the project's mandate and in compliance with the highest financial and ethical standards. In addition, the directors must present a public defence at the conclusion of their projects.

Stjepan Risović

Technology Council, Ministry of Science, Education and Sports, Trg Josipa Jurja Strossmayera 4,
10000 Zagreb, Republic of Croatia

Dwindling fish numbers already of concern in 1883

SIR — Jennifer Devine and colleagues in their Brief Communication “Deep-sea fishes qualify as endangered” (*Nature* **439**, 29; 2006) state: “At one time it was presumed from the vastness of the oceans that fishing would not drive species to extinction.” Indeed, the UK Royal Commission on the Sea Fisheries in 1866, whose officers included Thomas Henry Huxley, reported that fears of over-exploitation were unfounded. The commission recommended that existing laws regulating fishing grounds and closed seasons should be removed. But the rise in fish trade and reports about the scarcity of fish from all around the United Kingdom's coasts strengthened widespread concern among fishing communities and in scientific circles.

It all came to a head in 1883 at the International Fisheries Exhibition in London, a conference called to discuss commercial and scientific aspects of the fishing industry. In his inaugural address Huxley repeated the views of the royal commission by discounting reports of declines in fish catches. “With existing methods of fishing,” he said, “it is inconceivable that the great sea fisheries, such as those for cod, herring and mackerel, could ever be exhausted.”

Fortunately, there were some present who ventured to disagree. Their views were put forward by Edwin Ray Lankester.

“It is a mistake to suppose that the place of fish removed on a particular fishing ground is immediately taken by some grand total of fish, which are so numerous in comparison with man's depredations as to make his operations in this respect insignificant,” said Lankester. “If man removes a large proportion of these fish from the areas which they

inhabit, the natural balance is upset.”

Huxley and the royal commission did not foresee the advances in technology that accelerated depletion, particularly the move from sail to steam and then motor trawling. Even as late as 1919 there were influential British scientists, such as W. C. McIntosh, who denied that the sea's bounty could be exhausted by human activity (*Nature* **103**, 355–358 and 376–378; 1919, and Walter Garstang's response, *Nature* **104**, 48–49; 1919). But the earlier actions of their opponents ultimately stimulated the formation of several UK marine laboratories, such as those at Plymouth, Lowestoft and Millport, that conduct fish biology and fishery investigations to this day.

It is sobering to note that the concerns and problems facing people some 120 years ago are in some ways similar to those of scientists, industry and administrators today. Now, however, worldwide marine-fish populations are declining at an unprecedented rate, requiring greater international cooperation, research capacity and timely action.

David W. Sims, Alan J. Southward

Marine Biological Association of the UK,
The Laboratory, Citadel Hill,
Plymouth PL1 2PB, UK

Problems at plutonium lab need more than a quick fix

SIR — In addition to the concerns expressed in your Editorial “Enough, already” (*Nature* **438**, 712; 2005), about the planned increase in plutonium stored at Lawrence Livermore National Laboratory (LLNL), numerous other problems exist which do not lend themselves to ready fixes. LLNL is very near two earthquake faults, one of which — the Greenville fault — produced a 5.5-magnitude quake in 1980, causing extensive damage to lab buildings and opening a 120-metre discontinuous crack on the site.

LLNL is also a designated Superfund Site, identified by the Environmental Protection Agency for contaminants to be cleaned up. Local parks, gardens and yards have been shown to contain varied, sometimes unknown, levels of plutonium as the result of contaminated sewage sludge, according to a report by the California Department of Health Services in November 2002.

Tritium is also found on many sites in the Livermore valley: local wines contain concentrations of it well above background levels. During the past few years, these concentrations have decreased, but it is feared that the increase in use of tritium in the National Ignition Facility and other lab operations will reverse the trend.

Martha Priebat

3375 Norton Way 2, Pleasanton,
California 94566, USA

BOOKS & ARTS

A healthy dose of genetics

How much has genetics contributed to advances in medicine?

Moments of Truth in Genetic Medicine

by Susan Lindee

Johns Hopkins University Press: 2005.

288 pp. \$40

Uta Francke

The sequencing of the human genome is widely touted as a watershed event in biomedical science, and has brought an expectation that genetic information will revolutionize the practice of medicine. A reality check is in order, however. An examination from a historical viewpoint of how technical progress in medical genetics has led to clinical advances may modify our views of what 'genomic medicine' is likely to deliver.

In *Moments of Truth in Genetic Medicine*, Susan Lindee, a historian at the University of Pennsylvania, Philadelphia, focuses on 'case studies' in the history of human and medical genetics, emphasizing the family, social and political context of genetic disease. She examines the contributions of lay individuals and organizations to researchers' successes, and explores the connections between academic research, biotechnology, and financial and political motivations. She initially focuses on landmark developments between 1955 and 1975, but brings the story to the present day. She critically assesses the role of gene discoveries in finding cures and the potential for gene therapy. And she declares that the selective abortion of affected fetuses following prenatal genetic diagnosis is the "primary intervention of genomic medicine" to date. The more technically and financially demanding pre-implantation genetic diagnosis and embryo selection are not discussed.

The five case studies are well chosen to illustrate different aspects. 'Two Peas in a Pod' focuses on the use of twin research to determine the genetic component of human behaviour, exemplified by studies of alcoholism. 'Baby's Blood' explains in detail how the Guthrie blood-spot diagnostic test for phenylketonuria led to the establishment of screening programmes for newborns, with the aim of identifying affected homozygotes and instituting dietary treatment to prevent the development of mental retardation. Curiously, the implications of screening results for recurrence risk and carrier status of other family members are not mentioned. One wonders whether these genetic implications were even

made clear in the consent forms in the 1960s.

'Squashed Spiders' conveys the excitement about the discovery that having the wrong number of chromosomes can cause clinical syndromes. Karyotyping was the first widely used genetic diagnostic test for developmental disorders, and drew paediatricians, in particular, into the emerging field of medical genetics. Once the correct human chromosome number was established, an international effort to standardize the human karyotype was undertaken, even before most chromosomes could be unequivocally identified. Unfortunately, Lindee fails to explain correctly which chromosomes were individually identifiable and which were not. The international committee set up to standardize human cytogenetics subsequently developed numbering systems for chromosome bands revealed by different techniques.

'Provenance and the Pedigree' describes Victor McKusick's work with an inbred population in Pennsylvania: the Amish. The story of a rare, inherited form of dwarfism with extra fingers exemplifies the social, emotional

and technical aspects of collecting pedigree information. Lindee beautifully illustrates the personal relationship that develops between the researcher and the people who agree to divulge information about their family and medical history that is essential for the research to succeed. With the pedigree as the "bedrock tool of human molecular genetics", the people are not only research subjects but collaborators too. From the perspective of today's human geneticist, encumbered by increasing regulatory requirements and restrictions, it is amazing to see how easy-going and casual the interactions were when McKusick and his colleagues made their field trips into Amish country.

An ongoing contribution from McKusick is the collection of information on mendelian traits. First published in 1966 as *Mendelian Inheritance in Man*, this database is continually updated online, where it is known as Online Mendelian Inheritance in Man (OMIM). Lindee says the motivation for the creation of this catalogue was the collection of literature reports for the naming and classifying of

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

A seven-day-old baby is given the Guthrie test for detecting phenylketonuria.

clinical syndromes to guide genetic counselling. McKusick's major purpose, however, was to create a list of traits that could subsequently be mapped. In this book, the term 'mapping' has several meanings and is often used synonymously with 'sequencing'. The correct definition of mapping is either to place genes into a linear order (genetic mapping) or to assign them to chromosomal regions (physical mapping). The contribution of OMIM to gene mapping and the subsequent completion of the human genome sequence is immense.

In 'Jewish Genes', Lindee tells the story of familial dysautonomia, a rare recessive disorder characterized by lack of tears, autonomic dysfunction and emotional lability. Starting with the delineation of the phenotype and the development of a diagnostic skin test, she then discusses the mapping of the location of the disease gene and prenatal molecular testing by linkage analysis in affected families, and finally the identification of the mutant gene and its function. She allows the reader to enter the body of an affected individual and imagine living with the disease. It has highly variable clinical manifestations, including episodes

of fever and high blood pressure, impaired taste, vomiting, insensitivity to pain, undue reaction to anxiety and lack of motor coordination. The story of familial dysautonomia also focuses on the interactions between foundations that distribute research money, collected from affected families and their friends, and researchers who accept the funds and commit to work on the disease. The spectrum of interactions ranges from mutual support and respect to competitiveness and secretive behaviour.

These fascinating, well-written stories portray what it is like to work in human or medical genetics, both in the clinic and as a researcher. Nowadays, medical genetics is an established medical speciality, unlike 'genomic medicine', which is a poorly defined futuristic term. Students and researchers in genetics and other fields of biomedicine are likely to gain new perspectives on their own work by reading this book. And because everybody is a healthcare consumer, general readers may well identify with certain actors in these stories. ■ Uta Francke is in the Department of Genetics, Stanford University School of Medicine, Stanford, California 94305-5323, USA.

pressures that produced cognition and emotion in humans are shared by social carnivores, primates and birds, and he applies the principle of evolutionary continuity to suggest that similar outcomes may have arisen in the minds of other species. In addition to the more conventional argument that homologues of emotion are likely in non-human apes, he argues that convergent evolution may have produced emotions in species more distantly related to humans. This is thought-provoking material for anyone who wonders what the world is like for other species.

The book focuses mainly on Bekoff's work with dogs. Readers seeking broader surveys might also be interested in *Animal Social Complexity* edited by Frans de Waal and Peter Tyack (Harvard University Press, 2003), or *Why Men Won't Ask For Directions* by Richard C. Francis (Princeton University Press, 2003).

Bekoff's essays on the interactions between humans and other species raise important questions for those who study animals or who simply care about their conservation and well-being. The section on anthropogenic effects focuses on Bekoff's work with prairie dogs and domestic dogs. These papers are informative but for a broader view of the impact of humans on animals, the interested reader might also look at *Wildlife and Recreation* by R. L. Knight and K. J. Gutzwiller (Island Press, 1995).

The ethics of our interactions with other species provides the basis for some challenging discussions. Bekoff often concedes that these issues cannot be resolved, but he makes a compelling case that the questions are worth our attention. He argues that we should apply the precautionary principle to our interactions

In the minds of animals

Animal Passions and Beastly Virtues: Reflections on Redecorating Nature by Marc Bekoff

Temple University Press: 2005. 320 pp.
\$70.50 (hbk), \$26.95 (pbk)

Scott Creel

In T. H. White's *The Once and Future King* (G. P. Putnam, 1958), Merlin educates a young King Arthur by transmogrifying him into other creatures. Sometimes Arthur is an ant, and considers everything in the world only as 'done' or 'not done'. Sometimes he is a falcon, overcome in mid-sentence by a desire to kill his conversational partner. For many of us, the game of wondering 'What is it like to be a wildebeest?' never loses its charm. In *Animal Passions and Beastly Virtues*, Marc Bekoff explores this question in the context of cognitive ethology.

The book is really a collection of papers that Bekoff published between 1977 and 2004 on five main topics. The sections on social behavior in canids and on the functions of play fit neatly into conventional behavioural ecology. The quality of these original papers varies, from an excellent and influential paper on the social ecology of coyotes to a paper on scent-marking with a single subject, Bekoff's dog Jethro.

In three sections on animal minds, interactions between humans and animals, and the ethics of studying animals, Bekoff skates boldly on to thin ice at the edges of science and philosophy. In the section on emotions, cognition and animal selves, Bekoff argues that

many species have minds, emotions and a sense of self that bear recognizable similarities to those of humans. As well as providing case studies, Bekoff argues that invoking the existence of emotions or intentions often provides the most parsimonious explanation of behaviour. He also notes that the selection

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

A caring embrace? Marc Bekoff explores the impact of humans on prairie dogs.

with animals. In conservation biology, we should err on the side of caution when forced to make recommendations based on imperfect information. He makes a thoughtful case that because we cannot be sure that animals are not conscious, emotional individuals, it is wise to treat them with dignity and tread lightly. We need to be reminded of this point.

In short, this is a thought-provoking book that forces the reader to consider issues that are important but are often left at the fringes of our work. I found a lot with which I agreed, and a lot with which I disagreed. Examples of each might be revealing.

First, two disagreements. Bekoff argues that "classical definitions cannot be given for key terms in cognitive ethology, but it is not neces-

sary to give them in order to have a viable field of research". On this issue, I agree with Donald Griffin, who in his book *The Question of Animal Awareness* (Rockefeller University Press, 1976) took pains to state at the outset "this is what I mean by mental experiences" and went on to define 'mind', 'awareness', 'intention' and 'consciousness'. In my view, one cannot test a falsifiable hypothesis that a species has intention without first clearly defining what intention is. Bekoff also supports Stephen Jay Gould's assertion that "the plural of anecdote is data". I agree that anecdotes can offer important insights, but anecdotes are collected and disseminated selectively, whereas data are collected systematically or randomly. The plural of anecdote is not 'data', but 'essay'.

And here are two agreements. "We will never learn about animal morality if we close the door on the possibility that it exists. It is still far too early to draw the conclusion that human morality is different in kind from animal morality and walk away in victory." And: "While ignorance may be bliss, ignoring questions about our ethical responsibilities to animals compromises not only their lives and our integrity, but also the quality of scientific research". Marc Bekoff, by forcing us to consider animal minds and our ethical obligations to animals, is pushing the field of behaviour in interesting directions.

Scott Creel is in the Department of Ecology, Montana State University, Bozeman, Montana 59717, USA.

A ship, gradually lowering sail

The Long History of Old Age

edited by Pat Thane

Thames & Hudson/Oxford University Press: 2005. 320 pp. £25/\$49.95

Douwe Draaisma

Even when advanced in years, Lady Sarah Cowper (1644–1720) was still a keen observer, especially of other women. After meeting with "Lady W." she confided to her diary that W. had "rent her face with painting. She is at least as old as I am and hugely infirm yet affects the follies and aires of youth, displays her breasts and ears, adorns both with sparkling gems while her eyes look dead, skin rivell'd, cheeks sunk, shaking head, trembling hands." Lady Sarah strongly felt that one should act one's age. She herself was perhaps 'elderly', but definitely not 'old'. She admitted to being old only when she was in her mid-sixties, and then it was related to diminishing eyesight and the loss of her faculties for reading and writing, rather than to her age as such.

Lady Sarah's observations are discussed in *The Long History of Old Age*, a delightfully written and illustrated monograph edited by Pat Thane, who has herself a long history of writing on old age. Defining 'old' in terms of infirmities and physical weakness, as Lady Sarah did, turns old age into a category with diffuse boundaries — certainly more diffuse than they are today, when in many Western countries 65 is the age of mandatory retirement. For most of the history of old age, pensions or annuities never coincided with retirement from work, and most people laboured as long as their health permitted.

The team of historians headed by Thane sets the record straight on many popular conceptions of old age. It is not true that before the twentieth century hardly anyone reached old age. Life expectancy was considerably lower, but this was due to high infant and child mortality rates. Those who reached adulthood had a good chance of living into

their sixties. It is equally untrue that older people are less respected today than they used to be, or that it was common for children to welcome their ageing, sickly parents into their own household.

In her introductory chapter, Thane provides evidence that the belief that the elderly were treated with more respect in the past is itself very old. As for two or three generations living together in one happy household, demographic statistics tell a grimmer tale. In the ancient Greek and Roman worlds, a ten-year-old would have only a 50% chance of having any of his grandparents alive. And if you lived to be 60 in the eighteenth century, you had only a one-in-three chance of having at least one surviving child. Given the average age of marriage and the infant mortality rate in the

nineteenth century, a three-generation family was probably rarer than a four-generation family today. When parents lived with the family of their children, it was often for reasons of need and poverty. Then, as today, the elderly liked to keep their independence as long as was reasonably possible.

What this book does superbly is the wedding of demographic information, scarce as this often is, to evidence taken from the visual arts, literature, philosophy and personal documents such as letters and diaries. Some of the finest parts of the story are in the illustrations. The captions are perceptive, brief essays in themselves, that are obviously written by an (unspecified) author with an eye trained in the history of art.

Considering the span — from antiquity to



The downward path: a seventeenth-century view of a man's progression through life.

MUSEUM KURHAUS KLEVE

the twentieth century — and the diversity of sources, it is only natural that the overall image of old age is one of variety. In the fourteenth century, when Dante compared the final stage of life with a ship gradually lowering sail before entering harbour — an image of tranquillity and acceptance — miniaturists painted old people to symbolize vices such as pride, sloth and avarice, reserving elegant maidens for the virtues of mercy and charity. Old and ugly were often felt to be natural companions, as were young and beautiful. What strikes me in the various depictions of old age, both literary and visual, is the prevalence of mockery and hostility. The old were sometimes viewed as experienced and wise, offering advice superior to the “beardless counsel” of the young, but more often as distrustful, garrulous,

obscene (if impotent) and miserly, sometimes all at once.

For a long time most old people were poor as well. In particular, day labourers and piece-workers earned less and less as their strength diminished. Often they were dealt the more menial jobs. There was no such thing as official demotion; nature took care of that by declining physical power. Under these conditions it makes sense that the grandparent spoiling the child only appears at the end of the seventeenth century, and then only in the leisured classes.

In this long history of old age the second half of the twentieth century can be seen to bring several paradoxes. Today there are many professions where the physical reasons for retirement have disappeared, yet retirement is still mandatory.

‘Old’ bodies are healthier and better conserved than ever — a 70-year-old looks perhaps like someone aged 60 at the beginning of the twentieth century — yet many older people seek cosmetic surgery to look younger or engage in activities associated with youth. It is still as if something isn’t quite right about being old and acting your age. Perhaps this resistance to ageing shows the tenacity of age-old stereotypes equating ‘old’ with ‘obsolete’ and ‘infirm’. What’s worse, these stereotypes may also reflect, as Lady Sarah Cowper’s comments on Lady W. testify, the opinions of those who are old themselves — well, elderly. ■

Douwe Draaisma is in the Department of Theory and History of Psychology, State University of Groningen, PO Box 72, 9700 AB Groningen, The Netherlands.

Nuclear-powered image

A short film playing with concepts of quantum physics makes it big.

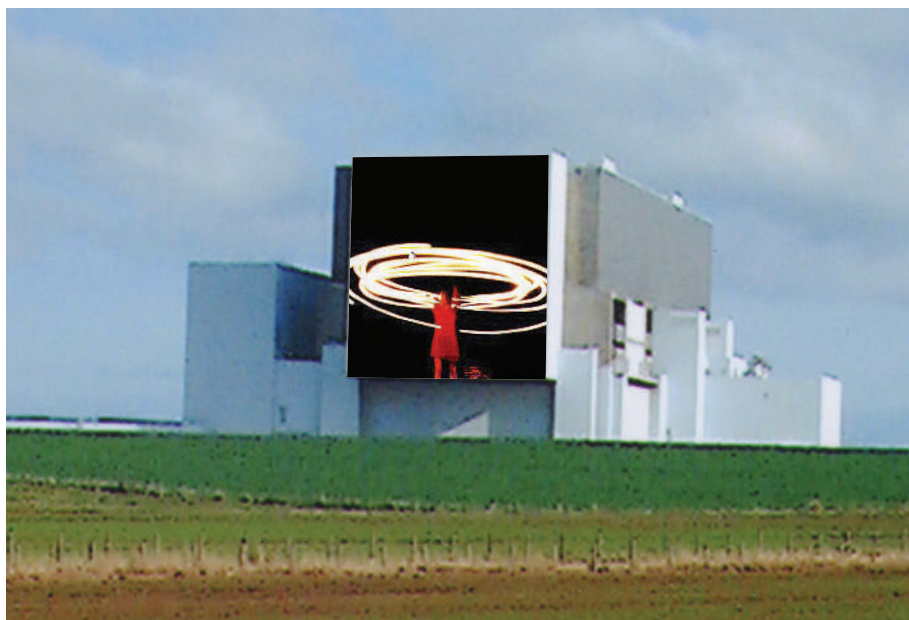
Alison Abbott

In Ken McMullen’s six-minute video loop *Lumen de Lumine*, filmed in an abandoned particle-accelerator tunnel at CERN, the European particle-physics laboratory in Geneva, a dancer in a cadmium-red dress swings a light bulb above her head, slowly letting out its flex.

At first she controls the bulb’s movement, but as the flex lengthens, the bulb’s increasing momentum is transferred to her body, which sways rhythmically to maintain balance. The bulb is the only source of light in the underground cavern, and when it moves behind her the woman is cast from light into darkness. The loop closes when she draws in the flex, puts down the bulb and the light is switched off.

On 9 February, in a preview of what it is hoped will be Europe’s largest ever public art installation, the lonely figure will be projected on to the giant wall of the Torness nuclear power station in Scotland. The loop will play continuously through the night. The promoter, Scottish arts impresario Ricky DeMarco, is looking for funding for a year-long run.

The film was created for the ‘Signatures of the Invisible’ exhibition in 2000, consisting of artworks produced in a collaboration between CERN physicists and European artists (see *Nature* **410**, 414; 2001). It has been remastered for this new setting, which adds its own dimensions to McMullen’s reflections on quantum physics.



K. MCMULLEN

As the light shuts off at the end of the loop, the woman whispers “Sein oder nicht sein, das ist die Frage”. The opening line of Hamlet’s most famous soliloquy is given in German, to acknowledge the German scientists who developed quantum theory. The film’s seemingly simple balletic sequence mirrors the contradictory elements of the quantum world, where the precise position of subatomic particles can never be known for certain. Does the figure dance in the light or in the dark? Is her dress red or black? Neither; both. The nuclear power station represents another duality by reminding us that we can split the atom for good, or for evil.

Torness power station is positioned within sight of the main road and rail links between

Edinburgh and London, used by some 16 million travellers each year. The travellers will never see the same image twice. Its clarity will change according to the weather and light conditions, and the image will also depend on the position from which it is viewed — a reference to Einstein’s relativity theory, where length and time change according to speed.

A discussion of some of the issues of ‘Signatures of the Invisible’ between artist and poet John Berger and CERN physicists is available on Ken McMullen’s DVD *Art, Poetry and Particle Physics*, part of his *Pioneers in Art and Science* series.

Alison Abbott is *Nature*’s senior European correspondent.

NEWS & VIEWS

PALAEONTOLOGY

A Jurassic tyrant is crowned

Thomas R. Holtz Jr

A newly discovered species of dinosaur, characterized by a wonderfully elaborate head crest, is the oldest known member of the lineage that culminated in *Tyrannosaurus rex* more than 90 million years later.

Dinosaur research — and indeed a whole swathe of palaeontology — has been revolutionized by recent discoveries in China. Most famous are the feathered dinosaurs, early flowering plants, various mammals and other spectacular fossils from the 128- to 110-million-year-old lake deposits in Liaoning Province in the northeast¹. Less well studied, however, are the many-coloured badlands of Xinjiang on the far western side of the country. These rocks contain fossils dating to the beginning of the Late Jurassic epoch, roughly 161 million to 156 million years ago.

From the Wucuiwan locality in Xinjiang come fossils of a new carnivorous dinosaur, described by Xu Xing and colleagues on page 715 of this issue². A pair of skeletons show that *Guanlong wucaii* ('crowned dragon of the five-coloured rocks'; Fig. 1) was a 3-metre-long predator. Details of the anatomy of this relatively small species indicate that it is the most ancient known member of the line that led to *Tyrannosaurus rex* and its giant kin, the Tyrannosauridae or 'tyrant dinosaurs' (Fig. 2).

Tyrannosaurids were the dominant group of predators in eastern and central Asia and North America during the last 20 million years of the Late Cretaceous epoch. Their morphology (enlarged skulls with enormous, robust teeth; highly reduced arms ending in two-fingered hands; and elongated hindlimbs), and above all their great size (9–13 metres long for the most completely known species), have made them among the most recognizable of fossil groups³. This distinctiveness, as well as their relatively rich fossil record, both in completeness of skeletons and numbers of individuals, has made the tyrannosaurids the subject of numerous palaeobiological studies^{4–6}.

Unfortunately, the uniqueness of the Tyrannosauridae has obscured their origin within the larger evolutionary tree of Theropoda — the clade (group) of carnivorous dinosaurs, including birds. The evolution of the distinctive adaptations present in the better-preserved Late Cretaceous forms such as *Tyrannosaurus*, *Gorgosaurus* and *Tarbosaurus* has transformed their skulls, limbs and vertebrae, thereby 'overwriting' much of the anatomical traces of their ancestry.

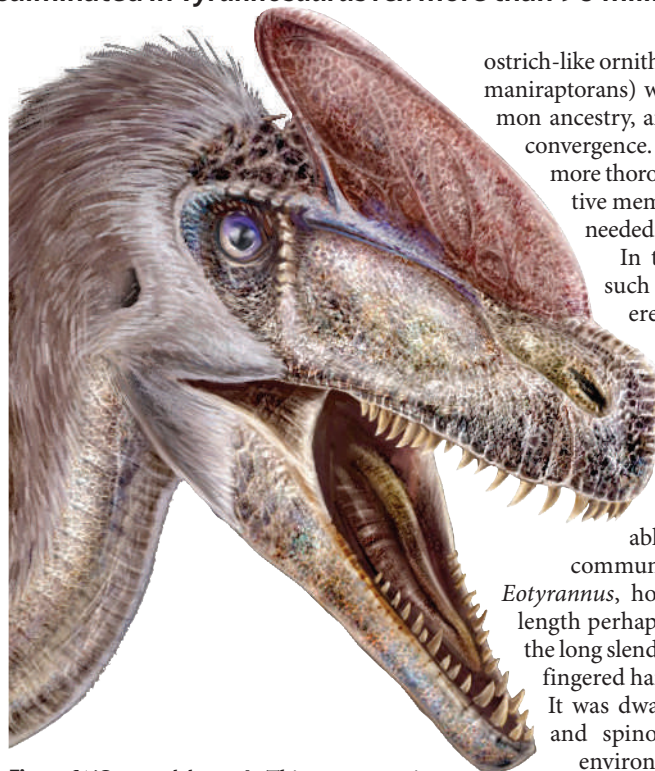


Figure 1 | 'Crowned dragon'. This reconstruction of the head of *Guanlong wucaii* is based on Xu and colleagues' data², the most notable feature being the crest, which may have been used in visual display. (Drawing by Zhongda Zhang, reproduced by courtesy of the Institute of Vertebrate Paleontology and Paleoanthropology, Beijing.)

For most of the twentieth century, palaeontologists followed H. F. Osborn⁷ in regarding the tyrannosaurids as the last descendants of Jurassic and Early Cretaceous carnosaurian giants, such as *Allosaurus* and *Acrocanthosaurus*. But in the 1990s, an alternative hypothesis⁸ that tyrannosaurids arose among the swift small-bodied coelurosaurs was ultimately supported by analyses⁹ that invoked the principles of cladistics. (Because cladistics provides testable hypotheses, this is now accepted as the best approach to evolutionary analysis.)

However, as long as older members of the lineage leading to the Tyrannosauridae proper were unknown, it remained questionable as to which anatomical transformations shared by tyrant dinosaurs and the various other coelurosaur groups (tiny compsognathids,

ostrich-like ornithomimosaurs, and feathered maniraptorans) were present owing to common ancestry, and because of evolutionary convergence. To address this question more thoroughly, fossils of more primitive members of the tyrant line were needed.

In the past few years, several such fossils have been discovered^{3,10,11}. Some, such as the Late Cretaceous *Appalachiosaurus* (Fig. 2, overleaf), were very close to true Tyrannosauridae in terms of size and anatomy, and, like their more specialized relatives, were probably the top carnivores of their communities¹⁰. Early Cretaceous *Eotyrannus*, however, was smaller (adult length perhaps 4.5 metres), and retained the long slender arms and grasping three-fingered hands of typical coelurosaurs. It was dwarfed by other (allosauroid and spinosauroid) predators in its environment, and probably played a secondary predatory role to its larger distant kin. In 2004, Xu Xing and colleagues described the Early Cretaceous *Dilong*¹¹, a 1.5-metre-long primitive carnivore with a covering of simple fuzzy 'protofeathers'. Until now, these dinosaurs represented the best record of early members of Tyrannosauroidae, a group composed of the giant two-fingered Tyrannosauridae and all of their closest kin.

The newly described species, *Guanlong*², reveals an even more primitive phase of tyrannosauroid evolution. Both skeletons are relatively complete, and in the smaller one, nearly the whole skeleton is still articulated, making *Guanlong* the best-preserved tyrannosauroid outside the Tyrannosauridae. Its adult length of 3 metres was smaller than all later tyrannosauroids save for *Dilong*. Like *Eotyrannus*, it was probably a secondary predator in its ecosystem, the primary predator apparently being the primitive carnosaur *Monolophosaurus*¹¹.

Guanlong combines characteristic tyrannosauroid traits (such as U-shaped teeth in the

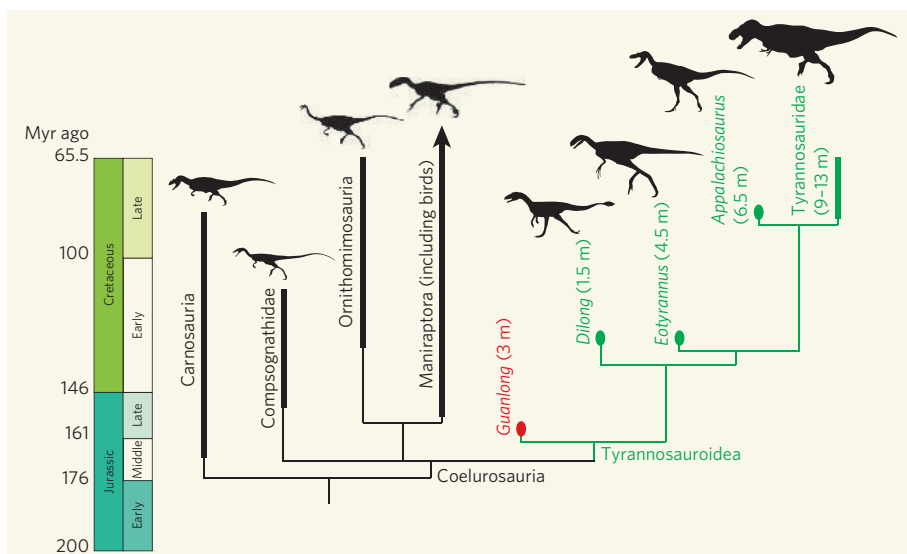


Figure 2 | *Guanlong wucaii* in evolutionary context. The bipedal and carnivorous theropod dinosaurs first appeared around 235 million years (Myr) ago in the Triassic period, later diverging into several different lines. In particular, during the Jurassic, one lineage (the Avetheropoda) divided into the Carnosauria and Coelurosauria, the latter including the Tyrannosauroidea³. Certain primitive features of the two *Guanlong* fossils² show that *Guanlong* lies at the base of the lineage of the tyrannosauroids, and close to the divergence between the major lineages of coelurosaurs. Numbers in parentheses indicate adult length.

premaxilla in the front of the upper jaw, and nasal bones fused into a single unit) with primitive characters (including relatively blade-like teeth along the sides of the jaws and a powerful three-fingered hand). Indeed, some aspects of the pelvis are unexpectedly primitive — more typical of Triassic-to-Middle Jurassic theropods than of the more advanced forms (although such features are present in at least one other coelurosaur¹²). The evolutionary analysis by Xu *et al.* incorporates the new information from *Guanlong*, and confirms that the Tyrannosauroidea arose towards the base of the Coelurosauria.

The most spectacular feature of *Guanlong* is its crest. All tyrannosauroids have some ornamentation along their nasal bones, be it a row of small hornlets as in *Appalachiosaurus* and *Alioramus*, a low ridge as in *Dilong*, or a roughened texture like all the others. But this newly discovered form has an especially impressive example — a tall, narrow projection with numerous hollow excavations. Narrow mid-line crests seem to have been the fashion for Middle and Late Jurassic theropods, as they are also present in *Ceratosaurus*, the primitive coelurosaur *Proceratosaurus*¹¹, and *Guanlong*'s possible predator *Monolophosaurus* (a previous record holder in terms of elaborate cranial crests among carnivorous dinosaurs). The fragile nature of these structures suggests that they served for visual signalling, and so for species recognition and mating displays, rather than as weapons. Indeed the fact that the smaller of the two *Guanlong* skeletons has a proportionately much smaller crest is consistent with the display being associated with sexual maturity.

Guanlong's presence at the beginning of the Late Jurassic, and the record of compsognathids and maniraptorans^{11,12} somewhat later

in the same epoch, indicate that the major divergences among the coelurosaurian dinosaurs had already occurred by 160 million years ago. Indeed, *Eshanosaurus*, a dinosaur

known only from an Early Jurassic lower jaw found in Yunnan, may represent a relatively advanced maniraptoran¹³. If so, the origin of all the primary lines of the coelurosaurs — including the tyrannosauroids — lies even further back. The new 'crowned dragon' of Xinjiang is simply the latest discovery on the trail leading back to the origin of the tyrant kings. ■

Thomas R. Holtz Jr is in the Department of Geology, University of Maryland, College Park, Maryland 20742, USA.

e-mail: tholtz@geol.umd.edu

1. Zhou, Z., Barrett, P. M. & Hilton, J. *Nature* **421**, 807–814 (2003).
2. Xu, X. *et al.* *Nature* **439**, 715–718 (2006).
3. Holtz, T. R. in *The Dinosauria* 2nd edn (eds Weishampel, D. B., Dodson, P. & Osmolska, H.) 111–136 (Univ. California Press, 2004).
4. Carr, T. D. *J. Vert. Paleontol.* **19**, 497–520 (1999).
5. Currie, P. J. *Can. J. Earth Sci.* **40**, 651–665 (2003).
6. Snively, E., Russell, A. P. & Powell, G. L. *Zool. J. Linn. Soc.* **142**, 525–553 (2004).
7. Osborn, H. F. *Bull. Am. Mus. Nat. Hist.* **22**, 281–296 (1906).
8. Matthew, W. D. & Brown, B. *Bull. Am. Mus. Nat. Hist.* **46**, 367–385 (1922).
9. Holtz, T. R. *J. Paleontol.* **68**, 1100–1117 (1994).
10. Carr, T. D., Williamson, T. E. & Schwimmer, D. R. *J. Vert. Paleontol.* **25**, 119–143 (2005).
11. Xu, X. *et al.* *Nature* **431**, 680–684 (2004).
12. Holtz, T. R., Molnar, R. E. & Currie, P. J. in *The Dinosauria* 2nd edn (eds Weishampel, D. B., Dodson, P. & Osmolska, H.) 71–110 (Univ. California Press, 2004).
13. Hartman, S., Lovelace, D. & Wahl, D. *J. Vert. Paleontol.* **25** (Suppl. to No. 3), 67A–68A (2005).

CHEMICAL BIOLOGY

Aptamers in nanoland

Michael Famulok and Günter Mayer

Chameleon-like nanoparticles of gold can be used to indicate the presence of various biomolecules. Adding aptamers — DNA strands that bind only to specific molecules — to the mix opens up further possibilities.

„Zum Golde drängt, am Golde hängt doch alles“ — “Towards gold throng all, to gold cling all”. This sentiment, from Goethe's masterpiece *Faust*, reflects mankind's fascination with this metal. But the lustre that makes gold so attractive in its bulk state changes entirely at the nanoscale. Gold particles of 10–100 nanometres possess optical properties that change according to their configuration: separated particles appear red in colour, and aggregated particles appear blue.

Juewen Liu and Yi Lu¹ exploit this chameleon-like nature of gold nanoparticles to create a colorimetric indicator for the detection of specific biomolecules. They publish their results in the journal *Angewandte Chemie International Edition*. The principle of the technique is not new: nanoparticles were first used as colour indicators to detect certain DNA sequences a decade ago^{2,3}, and have since become an important tool in bionanotechnology⁴. But Liu and Lu take the established principle

one step further by adding in the unique features of aptamers — short, single-stranded DNA sequences that recognize specific molecules and bind to them — to create a fast, analyte-specific indicator.

The more conventional part of the new approach¹ is loosely based on the concept of nanoparticle aggregation through DNA hybridization, a process in which two complementary single DNA strands anneal to form a double strand. First, each separated — and therefore red — gold nanoparticle has several single-stranded DNA sequences attached to it by covalent bonds (Fig. 1a). Two strands attached to neighbouring nanoparticles are then bound to one another through a parallel DNA linker consisting of a series of nucleotides complementary to both strands. As several identical DNA strands are attached to each nanoparticle, many nanoparticles can be glued together in this way. An aggregate forms, with a characteristic blue-purple colour (Fig. 1b).

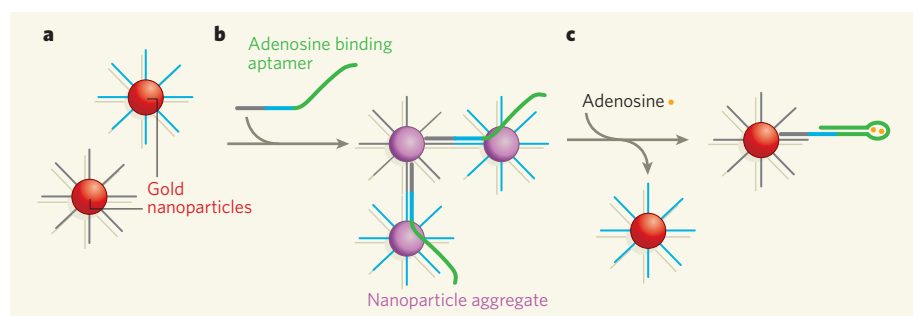


Figure 1 | Liu and Lu's aptamer-based detection strategy⁴. **a**, Gold nanoparticles, red when separated, are prepared with strands of DNA (radiating spokes) attached. **b**, A DNA aptamer (green) is modified so that one end is formed of consecutive sequences complementary to the DNA strands of two nanoparticles. The nanoparticles are thus linked by DNA hybridization, causing their colour to change to a blue-purple. **c**, The addition of adenosine (orange dots), the target molecule of the aptamer, causes dramatic conformational changes of the aptamer structure, which leads to the dissociation of the nanoparticle network. The colour of the solution changes back to red, a visual indication of the presence of adenosine.

The innovative part of Liu and Lu's strategy is that the DNA linker is itself an extension of a DNA aptamer that binds specifically to adenosine⁵, a representative small molecule. In the absence of adenosine, the aptamer behaves essentially like any short, linear DNA sequence, and does not adopt a well-defined three-dimensional structure. So any part of the aptamer is equally accessible for hybridization to a complementary sequence, and therefore to function as a linker.

In the presence of adenosine, however, the aptamer markedly changes its conformation (Fig. 1c). It folds to form well-defined binding pockets for adenosine molecules, exhibiting a three-dimensional structure stable enough to be elucidated by NMR spectroscopy⁶. The binding process also requires the participation of endmost nucleotides of the DNA linker to form the compact structure of the aptamer-adenosine complex. This can only happen if the gold nanoparticle right at the end of the aptamer is released, and so the whole network of aggregated gold nanoparticles disassembles. The colour of the ensemble therefore turns back to an intense red, and so acts as a visual indicator of the presence of adenosine.

This principle is not limited to adenosine detection, as Liu and Lu show with a neat second example¹. In this case, the colour change is triggered by a small molecule that might need to be detected in situations when elaborate analytical equipment is not readily available: cocaine. The authors found that, with slight adjustments of the appropriate aptamer, the dissociation of the DNA linker from one set of gold nanoparticles could be brought about upon cocaine binding to the aptamer. Using this system, the amount of cocaine present could be quantified in the range 50 to 500 micromoles per litre.

Both systems proved to be highly specific, exactly reflecting the specificity of the parent aptamers. The rapidity with which these sensors can detect analytes (a response is given within seconds), as well as the sensors' simple design and the fact that mere inspection is, in

principle, sufficient for semi-quantitative detection, are outstanding advantages of the technique, and ones that set it apart from previously developed aptamer-based detection approaches⁷.

The detection limit of the nanoparticle-based assay is correlated with the appropriate dissociation constant, a number that measures the strength of the small molecule's binding to the aptamer. As used by Liu and Lu¹, each nanoparticle has an average of 17 linkages, all of which need to be broken to disassemble the aggregate. So Liu and Lu's method is more energy-intensive than other detection principles, and so less sensitive. Improving the sensitivity of the technique could, however, easily be accomplished simply by reducing the number of linkages on each nanoparticle.

An alternative strategy for increasing detection sensitivity might be to apply the principles of Liu and Lu's approach to the distance-controlled assembly of another class of nanoscale materials — quantum dots. Quantum dots are nanocrystals (for example, zinc sulphide-capped cadmium selenide) that can be used as fluorescent probes in biodetection assays, including those for the detection of nucleic acids. Microbeads labelled with quantum dots of different sizes in controlled ratios can be made to exhibit a unique fluorescent signal for every DNA sequence⁸. By controlling the distance between two differently labelled quantum dots using an aptamer linker-and-target complex, one might be able to take advantage of a phenomenon known as fluorescence resonance energy transfer (FRET) to alter the emission wavelength of a quantum-dot probe, and so its colour. Writing in a recent edition of *ChemBioChem*, Andrew Ellington and colleagues⁹ have taken a first step in this direction, using an aptamer-labelled quantum dot, the fluorescence of which is reduced by an annealed fluorescent sequence. This sequence is then released in the presence of the analyte molecule to which the aptamer links⁹.

These aptamer-based tools^{1,9} could become a very useful addition to the nanostructure



50 YEARS AGO

"The 'Queen-Substance' of Honeybees and the Ovary-inhibiting Hormone of Crustaceans"
— In the course of work on the social organization of honeybee communities, it has been found that worker honeybees obtain a substance ('queen-substance') from their queens which, if obtained in sufficient quantity, inhibits development of their ovaries and the production of further queens. It will also, under experimental conditions, inhibit ovary development in worker ants (*Formica fusca*)... A similar hormone has been shown to inhibit ovary development in decapod crustaceans. These substances appear to be similar in several respects. Thus both are stable to heat and to acids but less to alkalis, soluble in acetone and alcohol, both are active, at least under certain conditions, when taken orally, and both serve to inhibit development of the ovary and related phenomena.

From *Nature* 11 February 1956.

100 YEARS AGO

"Result of War affected by Soldier's Stature" — The Japanese had an unquestionable advantage in the recent war as being smaller than the Russians; they were smaller targets for fire-arms. I wish to point out that it is possible to express this advantage quantitatively on the assumption... that bullets are, on average, uniformly distributed over the target presented by a man's body, also that a man presents a target proportional in area to the square of his height. The Anthropological Institute has kindly given me figures for the purpose; the average height of 2500 Japanese... was 1585 millimetres as compared with an average of 1642 millimetres for the average of 177,948 European Russian conscripts. The average Russian height thus exceeds that of the Japanese by about 3.47 per cent. The squares of the two average heights, representing, as I have said, the average targets offered by each to an enemy, differ therefore approximately by 7 per cent, so that the Russian fire was relatively ineffective to that extent.

John H. Twigg

From *Nature* 8 February 1906.

50 & 100 YEARS AGO

biodiagnostics workbox. Imagine that simple litmus tests for every analyte for which aptamers exist were available! To be able to look at a 'paper strip' to gain an instant idea of whether a compound is present and in what quantities would be worth more than its weight in gold.

Michael Famulok and Günter Mayer are at the Kekulé Institut für Organische Chemie und Biochemie, University of Bonn, Gerhard-Domagk-Strasse 1, 53121 Bonn, Germany. e-mail: m.famulok@uni-bonn.de

1. Liu, J. & Lu, Y. *Angew. Chem. Int. Edn* **45**, 90–94 (2006).
2. Alivisatos, A. P. *et al.* *Nature* **382**, 609–611 (1996).
3. Mirkin, C. A., Letsinger, R. L., Mucic, R. C. & Storhoff, J. J. *Nature* **382**, 607–609 (1996).
4. Rosi, N. L. & Mirkin, C. A. *Chem. Rev.* **105**, 1547–1562 (2005).
5. Huizenga, D. E. & Szostak, J. W. *Biochemistry* **34**, 656–665 (1995).
6. Lin, C. H. & Patel, D. J. *Chem. Biol.* **4**, 817–832 (1997).
7. Stojanovic, M. N. & Landry, D. W. *J. Am. Chem. Soc.* **124**, 9678–9679 (2002).
8. Han, M., Gao, X., Su, J. Z. & Nie, S. *Nature Biotechnol.* **19**, 631–635 (2001).
9. Levy, M., Cater, S. F. & Ellington, A. D. *ChemBioChem* **6**, 2163–2166 (2005).

IMMUNOLOGY

Exhausted T cells perk up

Matthew A. Williams and Michael J. Bevan

During persistent infections, the immune cells responsible for killing infected cells and maintaining inflammation gradually stop functioning, allowing the pathogen to thrive. But can this process be reversed?

There are many checks and balances in place to control the immune response to a pathogenic onslaught: too little of a reaction and the invader can kill the host or take up residence; too much and the immune system can itself start to damage tissues. On page 682 of this issue, Barber *et al.*¹ demonstrate that a receptor–ligand pair involved in controlling the immune response can be blocked by antibodies to revive dysfunctional T cells and clear a chronic viral infection.

Following many infections, immune cells that recognize antigens from the invading pathogen (CD8⁺ T cells) undergo a dramatic expansion in number and develop functions to deal with the intruders. These 'effector' T cells can kill infected cells and produce cytokine molecules, such as interferon- γ and tumour-necrosis factor- α , to promote local inflammation. The pathogen-specific T cells go through 15 or more cell divisions in five days, producing millions of effector cells that go on a search-and-destroy mission to clear the infection. These enormous numbers of effector cells cannot be maintained once the infection is gone, so most of them die. But they leave behind a pool of stable, long-lived memory T cells that can deal promptly, it is hoped, with a second exposure to the same pathogen (Fig. 1a). Some infections do not follow this textbook 'acute' course, however. The pathogen may find ways to avoid clearance, and the infection becomes persistent, or 'chronic'.

Lymphocytic choriomeningitis virus (LCMV) is a natural pathogen of mice, and most LCMV infections of healthy mice follow the acute course. However, there is a lab-derived variant of LCMV that overwhelms the immune response (either by

replicating faster or by infecting different host cells), resulting in a persistent infection with high levels of virus in the blood and many other tissues. Initially, mice infected with the lab-derived variant develop robust T-cell responses that are comparable to those induced by the parent strain. Then things start to deteriorate. Instead of clearing the virus within a week, the viral levels remain high, and

the effector cells gradually dwindle in number and lose the ability to kill cells and make cytokines. In immunology jargon, they become exhausted (Fig. 1b).

Barber *et al.*¹ noted that although the early, functional T cells in acute or chronic infections make a lot of a surface receptor protein known as PD-1, after clearance of the acute infection the resting memory cells lose expression of this marker. In chronically infected animals, however, the T cells retain high levels of PD-1 expression. In addition, one of the ligands for this receptor, PD-L1, is expressed at high levels on the surface of chronically infected cells. This receptor–ligand system has been postulated² to inhibit signalling through the T-cell antigen receptor — the receptor that triggers the response to a pathogen.

Spectacularly, the authors found that the decline in number and function of the effector T cells in chronically infected mice could be reversed by injecting an antibody that blocked the interaction between PD-1 and PD-L1 (Fig. 1c). They show that exhausted T cells that could not divide in response to antigen recovered this ability in the presence of the antibody. In addition, the T cells regained killer activity, secreted increased levels of interferon- γ and tumour-necrosis factor- α and — most importantly — rapidly reduced the viral load.

The question remains as to whether a short course of such blocking antibodies could have beneficial effects on persistent infections in people; for example, those caused by hepatitis virus, cytomegalovirus or HIV. Barber *et al.*¹ show that even in mice that lack CD4⁺ helper T cells, and therefore experience a much more pronounced course of LCMV infection and T-cell exhaustion, PD-L1 blockade can still revitalize the exhausted T cells and result in diminished virus levels. This is of particular significance for HIV infection, which is characterized by declining numbers of CD4⁺ T cells.

The first step in finding out whether this pathway can be exploited therapeutically will be to study the expression pattern of this receptor and its ligand in humans during a variety of chronic infections. But persistent antigen exposure also occurs in cancer, where continual presentation of antigens derived from abnormal 'self' proteins can lead to dysfunctional, tumour-specific T cells³. Notably, many tumours express PD-L1, so blocking this pathway may be beneficial in encouraging the immune system to recognize and kill off the abnormal cells.

Regardless of whether the PD-1–PD-L1 pathway is relevant in other models of chronic antigen presentation, Barber and colleagues' results demonstrate that the exhaustion of T cells during chronic infection can be reversed, raising the possibility of therapies designed to revive T-cell function. But first, the balance between regulatory and activation pathways that govern the T-cell responses during chronic infections

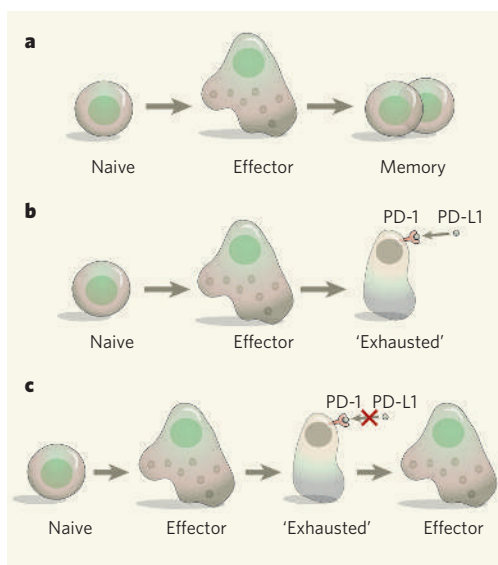


Figure 1 | Reviving exhausted T cells. **a**, In an acute infection, pathogen-specific T cells respond and develop into effector cells that help to eliminate the invader by killing infected cells and secreting cytokines. Once the infection is cleared, the remaining effector cells give rise to long-lived, protective memory cells. **b**, In a chronic infection, with a high antigen load, the effectors gradually lose function and the ability to proliferate. **c**, Barber *et al.*¹ show that in mice these 'exhausted' T cells can be revived by treatment with an antibody that blocks the interaction between the PD-1 receptor and its ligand PD-L1.

needs to be defined better, as there could well be hazards in interrupting such signals. For instance, mice in which the gene encoding PD-1 has been inactivated develop spontaneous autoimmunity⁴. Moreover, Barber *et al.* report that infecting mice that lack PD-L1 with the persistent virus does not lead to clearance but rapidly to death, probably because of an overreaction of the immune system. Another danger was highlighted when two mouse strains were mated; when the resulting pregnant mice were treated with a similar antibody against PD-L1, they aborted a large fraction of the fetuses⁵.

Although there were no harmful side effects when Barber *et al.* used antibodies to block the interaction between PD-1 and PD-L1, the

authors are conscious that autoimmune disease might develop after prolonged disruption of this regulatory system. Their results provide further evidence of the delicate balance governing immune regulation and activation. ■

Matthew A. Williams and Michael J. Bevan are in the Howard Hughes Medical Institute, and Department of Immunology, University of Washington, PO Box 357370, Seattle, Washington 98195-7370, USA.
e-mail: mbevan@u.washington.edu

1. Barber, D. L. *et al.* *Nature* **439**, 682–687 (2006).
2. Chemnitz, J. M. *et al.* *J. Immunol.* **173**, 945–954 (2004).
3. Blank, C., Gajewski, T. F. & Mackensen, A. *Cancer Immunol. Immunother.* **54**, 307–314 (2005).
4. Nishimura, H. *et al.* *Immunity* **11**, 141–151 (1999).
5. Guleria, I. *et al.* *J. Exp. Med.* **202**, 231–237 (2005).

EARTH SCIENCE

The rise and growth of Tibet

Andreas Mulch and C. Page Chamberlain

It is not difficult to be impressed by the grandeur of high mountainous regions, but it is difficult to reconstruct how the elevation of such regions evolved. A study of the Tibetan plateau does just that.

Is Everest now at its highest point, or was it once even loftier? What was the greatest height attained by the vast highlands of the Tibetan plateau, and when did this occur?

As described elsewhere in this issue by Rowley and Currie (page 677)¹, these questions can be tackled — if not yet answered definitively — by analysing the isotopic composition of ancient raindrops. With this approach, the authors show that Tibet continuously grew northward over millions of years in response to the thickening of Earth's crust associated with the collision of the Indian and Asian continental plates. The driving forces for this collision are generated deep in Earth's mantle. But the key to unravelling the uplifting history of the central Tibetan plateau is found in lake sediments on the plateau, some of which formed as long ago as 40 million years.

In these lakes and their surrounds, changes in the oxygen-isotope composition of surface water (which is controlled by regional climate and elevation) are recorded in sediments. Systematic variations in oxygen-isotope composition across the plateau reveal that spatially variable uplift of the plateau to 4,000 metres or more above sea level was intimately linked to the timing and rates of

convergence of India and Asia (Fig. 1). Rowley and Currie¹ estimate that uplift to 4,000 metres was initiated as long ago as 40 million to 50 million years, in the early stages of that convergence.

The evolution of mountain topography reflects the balance between tectonic forces in Earth's crust and upper mantle, and climatically

driven erosion at Earth's surface. Their relative role in controlling the rise of mountains remains unclear^{2,3}, but the problem can be approached by reconstructing the elevation history of large continental plateaux. Such studies can improve our understanding of the coupling between tectonics and long-term climate change. For example, the Tibetan plateau — the largest continental highland on Earth — is a major barrier to air flow in the atmosphere, and it has been suggested that uplift of the plateau triggered the onset of the Indian summer monsoon³.

The chemical fingerprint of rain and snow that precipitated on the Tibetan plateau is found in the oxygen-isotope composition of calcareous minerals in Tibetan lake sediments. This is expressed as $\delta^{18}\text{O}$, which is the $^{18}\text{O}/^{16}\text{O}$ ratio in the soils or sediments relative to the $^{18}\text{O}/^{16}\text{O}$ ratio in sea water. For example, the $\delta^{18}\text{O}$ of calcite formed in soils is related to the $\delta^{18}\text{O}$ of soilwater or groundwater by a temperature-dependent fractionation factor; so $\delta^{18}\text{O}$ in calcite formed in soil is a sensitive tracer of surface water that stems from precipitation. The underlying principle of oxygen-isotope altimetry is that water that precipitates as rain or snow becomes increasingly depleted in ^{18}O the higher up a mountain range that it falls⁴; systematic changes in $\delta^{18}\text{O}$ with elevation can then be used to infer relative elevation differences between the water source in the ocean and the elevation at which the rain or snow fell^{5,6}.

In the context of earlier isotopic studies^{6,7}, the low $\delta^{18}\text{O}$ values of carbonates found by Rowley and Currie¹ across the central Tibetan plateau reflect the south–north migration of high terrain in response to crustal thickening and the buoyant rise of Earth's surface. This finding agrees with the results of thermomechanical modelling for the growth and uplift history of the Tibetan plateau⁸.

It was proposed previously that the present stature of Tibet reflects tectonic processes involving thinning and delamination of Earth's crust and mantle during the Miocene (10 million to 8 million years ago)³. But it has since been suggested^{9,10} that the sequential rise and northward growth of Tibet started much earlier than that, in Eocene times, about 50 million years ago. The new results¹ support this idea that the Tibetan plateau is a long-standing topographic feature that arose from the collision between India and Asia, and is not the more recent product of other, deeper-seated processes. The results also show that stable isotopic data from sedimentary sequences do indeed record the long-term, high-elevation history of the plateau, and can provide absolute elevation constraints at

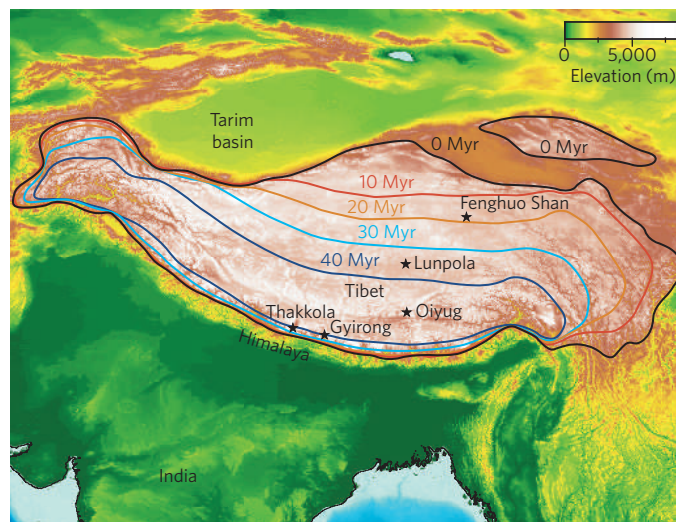


Figure 1 | Elevation history of the Tibetan plateau. According to Rowley and Currie's analysis¹, elevations of 4,000 metres or more were established as early as 40 million years (Myr) ago, with Tibet then growing northward over time. Solid lines show the development of the 4,000-metre contour as derived from oxygen-isotope data from the Thakkola, Gyirong, Oiyug, Lunpola and Fenghuo Shan sedimentary basins^{1,5–7}. The questions posed at the start of this article remain unanswered, however: the elevation history of Everest and the high Himalayan peaks has yet to be unravelled. (Figure courtesy D. B. Rowley.)

various times in the geological history of such a topographic feature. Moreover, taken in conjunction with independent estimates of palaeoelevation¹¹, this approach overcomes some of the limitations in deciphering the competing effects of climate and elevation change in the past.

Only slowly are we making progress in understanding how the interactions of surface uplift, bedrock erosion and sediment transport create dynamic feedbacks between the biosphere, the atmosphere and the crust and upper mantle. However, new techniques for estimating palaeoaltimetry that relate processes at various levels of Earth's crust, from the surface to deeper regions where high-temperature deformation causes rock flow, will allow us to develop a more dynamic

view of how mountain ranges change their shape and stature. Such techniques exploit elevation-dependent changes in concentration of carbon dioxide in the atmosphere¹², or the isotopic composition of surface waters that circulate deep in Earth's interior during the late tectonic evolution of mountain ranges¹³.

Deciphering the oxygen-isotope record in lake and soil deposits requires careful consideration of the competing effects of climate change and changes in surface elevation. The application of multi-proxy isotopic systems, which take account of both surface and deeper Earth environments, can complement such studies and greatly enhance our predictive capabilities in such a task.

Andreas Mulch and C. Page Chamberlain are in the Department of Geological and Environmental

Sciences, Stanford University, Stanford, California 94305, USA.
e-mail: mulch@pangea.stanford.edu

1. Rowley, D. B. & Currie, B. S. *Nature* **439**, 677–681 (2006).
2. Molnar, P. & England, P. *Nature* **346**, 29–34 (1990).
3. Molnar, P., England, P. & Martinod, J. *Rev. Geophys.* **31**, 357–396 (1993).
4. Dansgaard, W. *Tellus* **16**, 436–468 (1964).
5. Rowley, D. B., Pierrehumbert, R. T. & Currie, B. S. *Earth Planet. Sci. Lett.* **188**, 253–268 (2001).
6. Garzione, C. N., Quade, J., DeCelles, P. G. & English, N. B. *Earth Planet. Sci. Lett.* **183**, 215–229 (2000).
7. Currie, B. S., Rowley, D. B. & Tabor, N. J. *Geology* **33**, 181–184 (2005).
8. Beaumont, C., Jamieson, R. A., Nguyen, M. H. & Medvedev, S. J. *Geophys. Res.* **109**, B06406 (2004).
9. Tapponnier, P. et al. *Science* **294**, 1671–1677 (2001).
10. Graham, S. A. et al. *Am. J. Sci.* **305**, 101–118 (2005).
11. Spicer, R. A. et al. *Nature* **421**, 622–624 (2003).
12. McElwain, J. C. *Geology* **32**, 1017–1020 (2004).
13. Mulch, A. et al. *Geology* **32**, 525–528 (2004).

SEMICONDUCTOR PHYSICS

Transport news

John J. Boland

Conventionally, conduction in silicon is enhanced by doping — adding impurities that change the material's electronic structure. But exploiting surface effects in thin silicon films may offer yet other opportunities.

Silicon-based electronics continues to obey the dictum known as Moore's law: that the density of transistors on an integrated circuit — a rough measure of the attainable processing power — doubles about every 18 months. As the size of the smallest features of a device approaches the nanoscale, the electronic properties of the constituent materials are increasingly affected by the surrounding surfaces and interfaces. This can sometimes have deleterious effects on the transport of charge carriers — and so on device performance. On page 703 of this issue, Pengpeng Zhang and colleagues¹ show how surface effects can in fact be used to control conduction in silicon membranes. Their results potentially provide a new route to nanoscale materials and devices.

The continued drive towards higher device densities on silicon chips requires not only lithographical techniques that can define ever-smaller device features, but also the ability to manipulate precisely the Fermi energy of electrons in silicon, conventionally achieved by the process of doping (Box 1). But surfaces and interfaces are known to affect the electronic properties of silicon too. For instance, electrons from surface atoms can establish themselves within silicon's band gap, in which there are normally no allowed electronic states. There they can trap or deplete charge carriers, and so affect the position of the Fermi energy. Having the correct surface passivation chemistry — that is, the right sort of chemical bonds on the surface to eliminate mid-gap states — minimizes the effect of these rogue states. For

silicon, an exceptionally low density of interface traps is achieved by an interface with an insulating layer of silicon oxide, SiO₂. In this case, provided the dopant level is high enough, at most only a few silicon layers adjacent to the interface are depleted of carriers. The position of the Fermi energy is still controlled by the dopant concentration in the bulk of the material.

This interplay between doping and interface states is a fundamental aspect of device physics, and becomes even more important

when small structures are involved². Then, the lower total number of atoms, the increasing fraction of surface and interface atoms and the dopant population combine to provide an interesting mix. One possible consequence is fluctuation in the number of dopant atoms incorporated into the silicon, resulting in unacceptable variations in device performance across the wafer.

A possible solution to this problem is the development of high-resolution techniques that enable dopant ions to be implanted into the semiconductor material one by one³. But even if the dopant density could be controlled so precisely, the nanoscale membranes would eventually become so thin that even the low density of traps afforded by the SiO₂ interface would deplete the device of carriers, rendering it unworkable. Such effects have long been regarded as a fundamental limitation on the minimum size and thickness of silicon-based devices.

Zhang and colleagues¹ use a combination of

Box 1 Bands, holes and flows

In any solid material, the atomic orbitals that determine the properties of the electrons in an atom combine to form bands of allowed energy levels. Electrons fill these levels two-by-two following rules laid down by quantum mechanics. The energy up to which the allowed energy levels are filled is known as the Fermi energy. If this energy lies within a band of allowed energy levels, the material is a conductor; if it lies in a gap between allowed energy levels, the material is a semiconductor or insulator.

In a semiconductor, the gap between the completely filled 'valence' band and the completely empty 'conduction' band is small (about 1 electronvolt). So, whereas at low temperatures these materials are insulators, at normal temperatures a small population of electrons can be excited thermally from the valence band to the conduction band. The electrons create a negatively charged flow of current in the conduction band;

equally, the electron deficiencies they leave behind — known as 'holes' — give rise to what is in effect a positively charged flow of current in the valence band.

The number of electrons or holes can be tailored by adding so-called dopant atoms to the semiconductor. In a silicon crystal, with its four valence electrons per atom, this doping may be achieved in one of two ways. First, by introducing atoms of an element such as phosphorus, antimony or arsenic that has a fifth electron (n-type doping). These extra electrons will occupy energy states close to the conduction band, into which they are easily excited to contribute to conduction. Second, doping may be effected by introducing an impurity possessing only three electrons, such as boron, aluminium or gallium (p-type doping). This creates energy states close to the valence band, into which electrons can jump, leaving behind holes in the valence band.

J.J.B.

EVOLUTION

Memories of mammoths

If elephants never forget, the memories of mammoths need a little prompting. Nevertheless, inventive approaches to the extraction and sequencing of DNA from mammoths preserved in Siberian permafrost are allowing direct access to the deeper memories of elephant evolution.

There has been much debate about whether the woolly mammoth (*Mammuthus primigenius*) — that archetype of everything icy and Palaeolithic — was more closely related to the extant African or Asian elephant (*Loxodonta africanus* and *Elephas maximus*, respectively). Analysis of the complete mitochondrial genome of a mammoth by Hofreiter and colleagues (*Nature* **439**, 724–727; 2006) provides the answer: mammoths are more closely related

to Asian elephants, but only just.

Current wisdom has it that the lineages leading to mammoths and both extant elephant species diverged about 6 million years ago in Africa. The new data suggest that the African lineage split first, followed around 440,000 years later by the separation between Asian elephants and mammoths.

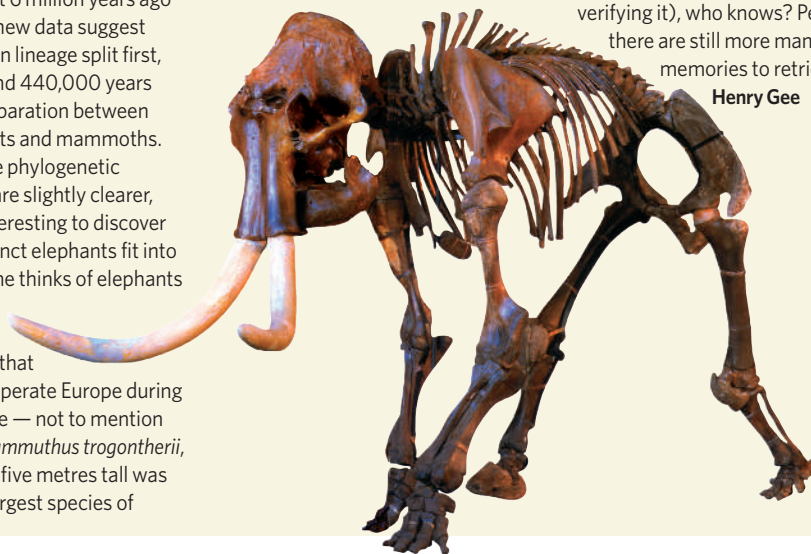
Now that the phylogenetic relationships are slightly clearer, it would be interesting to discover how other extinct elephants fit into the picture. One thinks of elephants such as *Anancus* and *Palaeoloxodon* that foraged in temperate Europe during the Pleistocene — not to mention the mighty *Mammuthus trogontherii*, which at up to five metres tall was possibly the largest species of

elephant ever, making the woolly mammoth look, if not dwarfed, then at least somewhat petite.

These questions may not be answered using ancient DNA, however. The permafrost environment seems to favour the preservation of ancient DNA in quantity, from mammoths as

well as other species, as shown by Poinar and colleagues (*Science* **311**, 392–394; 2006). But the likelihood of finding sufficiently informative DNA from species living outside the Arctic is almost certainly very much less. Yet, given the advances in sequencing ancient DNA (and, more importantly, verifying it), who knows? Perhaps there are still more mammoth memories to retrieve.

Henry Gee



STATE NAT. HIST. MUS., BRAUNSCHWEIG/PALEONT. INST., RUSS. ACAD. SCI.

scanning tunnelling microscopy and modelling to show that this conventional analysis cannot fully explain electron transport in nanoscale silicon films. Using silicon-on-oxide membranes, already employed as the basis of low-capacitance, fast-switching devices, they demonstrate that transport is feasible even with very thin membranes, and for doping levels at which the Si–SiO₂ interface effectively depletes all carriers within the semiconductor. Electronic conduction is seemingly, in this case, not controlled by dopants in the bulk of the solid.

So what does control electron transport in silicon membranes? The authors provide compelling arguments to implicate surface states that come about when the single dangling bond of two neighbouring silicon surface atoms combine to form so-called 2 × 1 dimers. This process results in filled and empty surface states that are very close in energy to the valence band of the bulk silicon, and, in the authors' model, generate holes that dominate charge transport (Fig. 1).

This behaviour is probably not unique to the 2 × 1 surface configuration, or even to silicon. In fact, any surface 'termination' that provides energy levels that are low enough to accept electrons from the valence band should suffice. In the case of silicon, this would obviate the need to work with the notoriously reactive 2 × 1 surface, which is impractical for any real device application. That would open up new avenues to the manufacture of nanoscale devices.

The surface states associated with the 2 × 1 surface can be eliminated by chemical

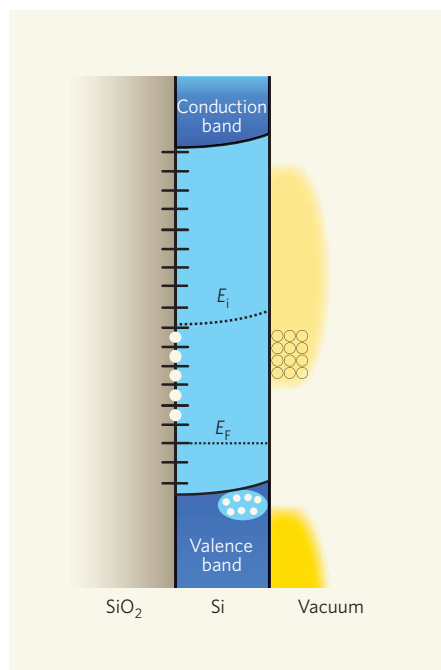


Figure 1 | The band structure of a silicon-on-insulator membrane. Trap states (short lines) at the silicon–silicon-dioxide interface in Zhang and colleagues' thin silicon membrane¹ depress the Fermi energy (E_F) of the membrane to below its intrinsic level (E_i), hindering conduction. But surface effects create allowed energy states (yellow) just above the valence band, into which electrons (blue circles) can be excited thermally, producing 'holes' (white circles) that can propagate through the valence band. This surface effect in silicon works equivalently to the doping effect in bulk silicon in enhancing the material's conduction properties. (Adapted from ref. 1.)

reaction: terminations such as hydrogen or halogens would suppress the surface states and prevent the generation of holes within the film. Such terminations could be patterned, providing a method of creating locally conducting and insulating regions of the substrate. But generating new surface states close to the energy gap is best achieved by tailoring the gap between the highest occupied and lowest unoccupied orbitals of molecular species attached to the surface⁴. This would in principle allow the generation of both hole- and electron-conducting regions in the film.

Although such surface-state models provide a potentially powerful route to control the electronic structure of nanoscale semiconductors, further studies will be needed to allow a detailed understanding of the transport mechanisms that result. Temperature-dependent conductivity measurements and high-resolution scanning tunnelling spectroscopy would be useful in this regard. Mechanism aside, Zhang and colleagues demonstrate¹ that, in the quest for ever-smaller electronic devices, surprises and opportunities abound.

John J. Boland is in the Centre for Research on Adaptive Nanostructures and Nanodevices (CRANN) and the School of Chemistry, Trinity College Dublin, Dublin 2, Ireland.
e-mail: jboland@tcd.ie

1. Zhang, P. *et al.* *Nature* **439**, 703–706 (2006).
2. Erwin, S. *et al.* *Nature* **436**, 91–94 (2005).
3. Shinada, T., Okamoto, S., Kobayashi, T. & Ohdomari, I. *Nature* **437**, 1128–1131 (2005).
4. Hamers, R. J. *et al.* *Acc. Chem. Res.* **33**, 617–624 (2000).

BRIEF COMMUNICATIONS

Krakatoa's signature persists in the ocean

This huge eruption slowed sea-level rise and ocean warming well into the following century.

We have analysed a suite of 12 state-of-the-art climate models and show that ocean warming and sea-level rise in the twentieth century were substantially reduced by the colossal eruption in 1883 of the volcano Krakatoa in the Sunda strait, Indonesia. Volcanically induced cooling of the ocean surface penetrated into deeper layers, where it persisted for decades after the event. This remarkable effect on oceanic thermal structure is longer lasting than has previously been suspected¹ and is sufficient to offset a large fraction of ocean warming and sea-level rise caused by anthropogenic influences.

We examined the latest series of coupled ocean–atmosphere model experiments that include time-varying external forcings (for example, changes in greenhouse gases, solar irradiance, sulphate aerosols and volcanic aerosols) for the period 1880–2000 (see supplementary information). These models differ in their physics, resolution, initialization and ocean–atmosphere coupling procedures, and have different combinations of external forcings. Uncertainties in both the applied forcings and in the model responses to them are therefore inherent in our investigation.

We compared the evolution of global ocean heat content over 1880–2000 in six models that include the effects of volcanic eruptions (V) with six that do not (see supplementary information). Observations (which are subject to uncertainties arising from incomplete space- and time-varying coverage^{2,3}) suggest that the heat content of the upper 3,000 m of the ocean increased at a rate of $0.33 \times 10^{22} \text{ J yr}^{-1}$ over 1955–98. This is in closer agreement with the V simulations (average: $0.2 \times 10^{22} \text{ J yr}^{-1}$; with 1 s.d.: $0.16 \times 10^{22} \text{ J yr}^{-1}$) than with the simulations that do not include V (average: $0.78 \times 10^{22} \text{ J yr}^{-1}$; with 1 s.d.: $0.25 \times 10^{22} \text{ J yr}^{-1}$).

An abrupt drop in heat content in the V simulations (Fig. 1a) follows the 1883 Krakatoa eruption, augmented by much smaller eruptions in 1886 and 1888. Volcanic aerosols scatter and absorb sunlight and result in a cold ocean-surface-temperature anomaly. This is gradually subducted into deeper layers^{1,4}, where it persists for decades (Fig. 1b). Although surface warming in the late twentieth century is apparent in all V simulations, a

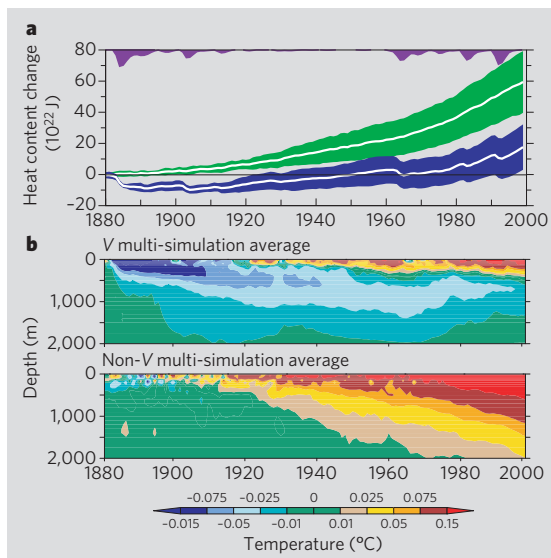


Figure 1 | Simulations with and without volcanic forcing (1880–2000). **a**, Change in global ocean heat content. Shading represents the ± 1 s.d. range of simulations with (blue) and without (green) volcanic forcing, V, about the corresponding multi-simulation means (white lines). Purple shading at the top is an estimate of changes in the amount of volcanic dust in the stratosphere, a measure of the reduction in sunlight reaching the Earth's surface (arbitrary scale). **b**, Global ocean-temperature anomalies as a function of depth for the mean of the simulations with and without V. The inter-simulation s.d. (not shown) decreases with depth, increases with time, and is generally larger for the V simulations. (For methods, see supplementary information.)

cold anomaly remains discernible at depth.

In spite of substantial differences in model formulations and applied external forcings (and, in particular, uncertainties in volcanic forcing in the pre-satellite era⁵; see supplementary information), the distinction between the simulations with and without V in Fig. 1 is striking. Although solar forcing is included only in the V simulations, its effect is minimal.

An oceanic response to the 1991 Pinatubo eruption, which was comparable to Krakatoa in terms of its radiative forcing, has been identified in satellite altimetry data¹. The simulated heat-content recovery after Pinatubo seems to occur much more rapidly than for Krakatoa (Fig. 1a). This disparity arises because the Pinatubo response is superimposed on a non-stationary background of large and increasing greenhouse-gas forcing. The heat-content effects of Pinatubo and other eruptions in the late twentieth century are offset by the observed warming of the upper ocean, which is primarily

due to anthropogenic influences⁶.

Ocean warming (or cooling) contributes to sea-level changes by thermal expansion (or contraction). Global mean thermal expansion is highly correlated with changes in heat content, and so comparisons of thermal expansion between the V and non-V simulations look much like Fig. 1a. Increases in thermal expansion at the end of the twentieth century (relative to 1882, the year before Krakatoa) are appreciably less for simulations with V (average: 1.7 cm; with 1 s.d.: 1.8 cm) than for the simulations without V (average: 6.3 cm; with 1 s.d.: 2.2 cm).

In model simulations, Krakatoa has long-lasting effects, offsetting a large fraction of the changes in ocean heat content and thermal expansion caused by twentieth-century anthropogenic influences. These results are robust to current uncertainties in climate models and in the historical forcings applied to them. Inclusion of volcanic forcing from Krakatoa (and, by implication, from even earlier eruptions) is important for a reliable simulation of historical increases in ocean heat content and sea-level change due to thermal expansion.

P. J. Gleckler*, T. M. L. Wigley†, B. D. Santer*, J. M. Gregory‡, S. K. AchutaRao*, K. E. Taylor*

*Program for Climate Model Diagnosis and

Intercomparison, Lawrence Livermore National Laboratory, Livermore, California 94550, USA
e-mail: pgleckler@llnl.gov

†National Center for Atmospheric Research, Boulder, Colorado 80307-3000, USA

‡Department of Meteorology, University of Reading, PO Box 243, Reading RG6 6BB, UK
§Met Office Hadley Centre, Exeter, Devon EX1 3PB, UK

1. Church, J. A., White, N. J. & Arblaster, J. M. *Nature* **438**, 74–77 (2005).
2. Levitus, S., Antonov, J. I. & Boyer, T. P. *Geophys. Res. Lett.* **32**, L02604 doi:10.1029/2004GL021592 (2005).
3. AchutaRao, K. M. et al. *J. Geophys. Res.* (in the press).
4. Delworth, T. L., Ramaswamy, V. & Stenchikov, G. L. *Geophys. Res. Lett.* **32**, L24709 (2005).
5. Robock, A. *Rev. Geophys.* **38**, 191–219 (2000).
6. Barnett, T. P. et al. *Science* **309**, 284–287 (2005).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Received 26 September 2005; accepted 22 December 2005.

doi:10.1038/439675a

Palaeo-altimetry of the late Eocene to Miocene Lunpola basin, central Tibet

David B. Rowley¹ & Brian S. Currie²

The elevation history of the Tibetan plateau provides direct insight into the tectonic processes associated with continent–continent collisions. Here we present oxygen-isotope-based estimates of the palaeo-altimetry of late Eocene and younger deposits of the Lunpola basin in the centre of the plateau, which indicate that the surface of Tibet has been at an elevation of more than 4 kilometres for at least the past 35 million years. We conclude that crustal, but not mantle, thickening models, combined with plate-kinematic solutions of India–Asia convergence, are compatible with palaeo-elevation estimates across the Tibetan plateau.

The elevation history of the Tibetan plateau is one of the few direct ways of discriminating among models for the modes of strain accommodation in the India–Asia continent–continent collision system. Data constraining the elevation history of the Himalaya and Tibet can have a critical role in determining the degree to which various mechanisms—including crustal subduction¹, crustal thickening^{2,3}, crust and mantle–lithosphere thickening with subsequent convective removal of the mantle lithosphere^{4–6}, as well as escape^{7,8} and extrusion⁹—have contributed to the high elevation and distribution of strain across the region. In this Article, we present new data from the Lunpola basin in central Tibet (Fig. 1) that constrain an ~40 Myr elevation history of this part of the India–Asia collisional system using oxygen-isotope palaeo-altimetry of Eocene–Oligocene and Miocene palaeosol carbonates and lacustrine limestones. We summarize existing palaeo-altimetric estimates in comparison with retrodictions (predictions about the past) made within the context of India–Asia convergence history since the onset of collision.

Regional setting of Lunpola basin

The Lunpola basin is primarily a Tertiary sedimentary basin situated in the central part of the Tibetan plateau at about 32° N, 89.75° E (Fig. 1). The basin is situated approximately half way between the Qaidam basin to the north and the Himalaya to the south. The basin axis strikes essentially east–west and has been variably deformed along basin-parallel thrust faults and associated folds. Erosion through these structures has exposed much of the succession at the surface. The Lunpola basin contains a thick stratigraphic succession that extends from the Palaeocene to the Pliocene¹⁰. The basin has been explored geophysically and through a series of wells¹¹ that has established the stratigraphic context and correlation of the sediments within the basin^{10,11}.

The Cenozoic strata of the Lunpola basin are more than 4,000 m thick, and consist of two primary stratigraphic units: the Palaeocene–Oligocene Niubao Formation, and the Miocene–Pliocene Dingqing (also referred to as the Dingqinghu) Formation. The age of these formations is based primarily on fossil ostracode and palynological assemblages^{10,12–14}. As part of our investigation, we sampled lacustrine and palaeosol carbonates from the middle and upper Niubao Formation, respectively, as well as lacustrine marls and limestones from the middle Dingqing Formation. In terms of age, the middle

Niubao Formation is classified as Eocene and the upper member is late Eocene to Oligocene, while the middle Dingqing Formation is Miocene in age^{10,12–14}.

Our sampling in the Niubao Formation was conducted in outcrops along the northern margin of the basin (Fig. 1 inset, location B). Here about 2.2 km of the middle and upper Niubao Formation are exposed in the north limb of a fault-propagation anticline in the hanging wall of the northern basin-margin thrust. At this location, the middle Niubao Formation consists of lacustrine mudstones and dolomitic marls, with minor beds of sandstone, algal dolostone, kerogenous shale, marl and micritic limestone. Because of our desire to make estimates of palaeometeoric precipitation based on oxygen isotopic composition of carbonate, we limited our sampling to the calcareous marls and micritic limestones preserved in the lower and middle parts of the middle Niubao Formation. At this location, the upper Niubao Formation consists of >400 m of fluvial/alluvial mudstone, sandstone and conglomerate that contains abundant intercalated calcic palaeosols near the top of the section. Pedogenic carbonate nodules from this interval were extensively sampled as part of this study.

We also sampled micritic lacustrine carbonates from the middle member of the Dingqing Formation in outcrops in the south-central parts of the basin (Fig. 1 inset, location A). The studied interval consists of mudstone, marl, micritic and oolitic limestone, sandstone and conglomerate. Some sandstones and siltstones contain symmetrically-rippled surfaces, as well as fossilized fish bones. Finer-grained lithologies were in places well laminated and contained fossil ostracodes and insects. Because of low structural dips and subdued topography, only 19 m of the unit is exposed at this location.

Isotope-based palaeo-altimetry

Rowley *et al.*¹⁵ described the physical and thermodynamic basis of a model of oxygen and hydrogen isotopes in precipitation in low latitude (<35°) orographic systems to estimate (palaeo-)elevations and potentially (palaeo-)hypsothermies of hydrologic systems. Ref. 15 presents a model of the relationship between elevation and $\Delta(\delta^{18}\text{O}_p)$, where $\Delta(\delta^{18}\text{O}_p)$ is the difference in isotopic composition of precipitation between that at sea level and that at elevation. Rowley *et al.*¹⁵ used data from the oldest intermontane basins within the Himalaya to demonstrate that the Himalaya have had their current stature for

¹Department of the Geophysical Sciences, 5734 S. Ellis Avenue, The University of Chicago, Chicago, Illinois 60637, USA. ²Department of Geology, 114 Shideler Hall, Miami University, Oxford, Ohio 45056, USA.

at least the last 8 to 10 Myr. Garzione *et al.*^{16,17}, using an empirical fit to surface water data, arrived at a similar conclusion.

Modern stream waters from Amdo to Lunpola

Comparison of the oxygen isotopic compositions of modern waters in southern Tibet with the model demonstrate that the modern system accords well with predictions^{15,18}. Small streams were sampled along the Amdo-Dongqiao-Lunpola road and within the Lunpola basin (Fig. 1). Elevations of the sampling sites range from 4,567 m to 4,718 m, based on GPS estimates. Maximum drainage basin heights are only a few hundred metres to a kilometre or so higher than the sample elevation, and hence there should not be a significant hypsometric effect in these samples. We compute predicted altitudes and palaeo-altitudes using equation (1) from ref. 18 (Fig. 2). Note that where samples are not directly measuring precipitation, as in surface and ground waters, we replace $\Delta(\delta^{18}\text{O}_p)$ with $\Delta(\delta^{18}\text{O}_w)$ to signify this, and to remind ourselves that $\Delta(\delta^{18}\text{O}_w)$ includes hypsometric effects of integrating precipitation in the drainage basin above the sample elevation. These Tibetan streams range from relatively unfractionated, with $\Delta(\delta^{18}\text{O}_w)$ of -3.2‰ , to highly fractionated, ranging up to -11.0‰ (Fig. 2). The majority of predicted elevations of these streams (Fig. 2) overlap within uncertainty with model expectations. It is important to recognize that modern waters do not plot above the model correlation line which would indicate fractionation by some combination of Himalayan orographic lifting and the effects of 'continentality'^{19,20}. In fact, other factors, not yet completely understood, reduce the altitude effect significantly for a number of these and other sample streams reported in the literature from central and northern Tibet²¹. If similar effects operated in the past, the estimated $\Delta(\delta^{18}\text{O}_w)$ would result in an underestimate of the palaeo-elevation in this region. Given our current data, we have yet to

encounter waters that are more depleted than predicted by the model¹⁵. We take this as a positive bias, as it leads to the expectation that the isotopes are much more likely to underestimate, rather than overestimate, (palaeo-)elevations.

Palaeo-elevation estimates of the Lunpola region

Palaeo-altitude estimates depend on $\Delta(\delta^{18}\text{O}_w)$. This requires the identification of low altitude reference sites. Ideally these would be coeval, unequivocally low elevation sites along the mean storm trajectories with which to compare data from the Lunpola basin. Unfortunately, there are no data from the region to the south of the appropriate ages for the sections described above. Instead we use modern observations to place bounds on likely compositions of low latitude and low altitude precipitation, and compare these with general circulation model-based estimates incorporating isotopes. In the modern world, there is little change in isotopic composition of low latitude, near sea level isotopic compositions of precipitation. Values of $\delta^{18}\text{O}_p$ of about $-3.6 \pm 1.6\text{‰}$ are found for average low latitude ($< \pm 35^\circ$) and low elevation stations ($< 100\text{ m}$), based on data released by the IAEA Global Network of Isotopes in Precipitation for years through to 2001²². Given these estimates, we adopt a value of $\delta^{18}\text{O}_p$ of $-6 \pm 1.6\text{‰}$ for the low elevation precipitation, comparable to the modern precipitation in New Delhi and more depleted than the average low latitude, low elevation, coastal value observed today. By adopting a more depleted low elevation value, we decrease our estimated palaeo-elevations.

The Dingqing Formation represents the youngest pre-Quaternary strata in the Lunpola basin. Our sampling focused on thin-bedded marls and fine grained limestones of the Miocene middle Dingqing Formation. Oxygen isotopic compositions of the lacustrine carbonate ($\delta^{18}\text{O}_c$) range from -1.3‰ to -14.6‰ (with respect to the Pee Dee belemnite (PDB) standard). It is important to note that the most widely ranging values in this section come from samples of thin (3–5 cm thick) marls that are only 50 cm apart. Figure 3, main panel, plots $\Delta(\delta^{18}\text{O}_w)$ versus predicted palaeo-altitudes of these Miocene samples. For each sample, we plot the mean $\Delta(\delta^{18}\text{O}_w)$ and the $\pm 2\sigma$ uncertainty in $\Delta(\delta^{18}\text{O}_w)$. This reported uncertainty reflects most

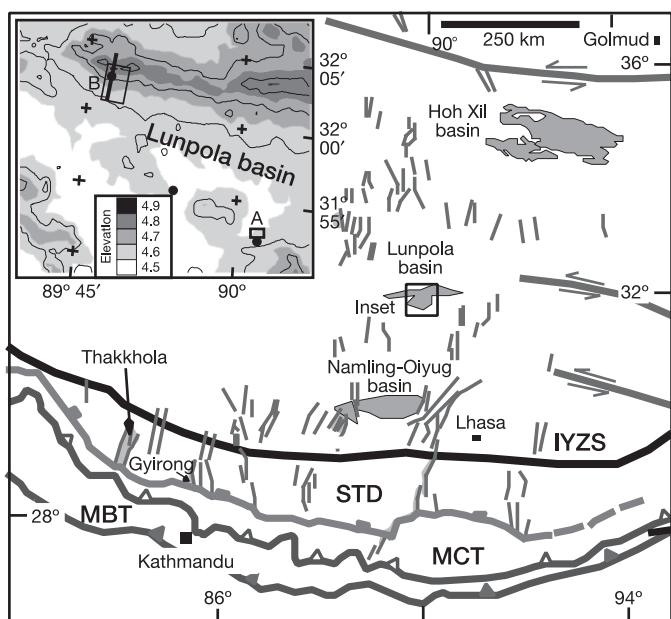


Figure 1 | Generalized tectonic map of the Himalaya-Tibet region, and details of the Lunpola basin. In the main figure are shown basins within the plateau for which palaeo-elevation estimates discussed in the text are derived. The location of the inset is shown as a box. Inset, map showing details of the Lunpola basin. Topographic contours are derived from Hydro1K Asian data; key shows elevation in km above sea level. Boxes labelled A and B are sites from which samples for palaeo-elevation measurements were collected for the Dingqing and Niubao, respectively. Heavy line is the location of the middle and upper Niubao section. Black circles are sites of collection of modern streams. Abbreviations: MBT, Main Boundary thrust; MCT, Main Central thrust; STD, South Tibetan detachment; IYTS, Indus-Yarlung Tsangpo suture.

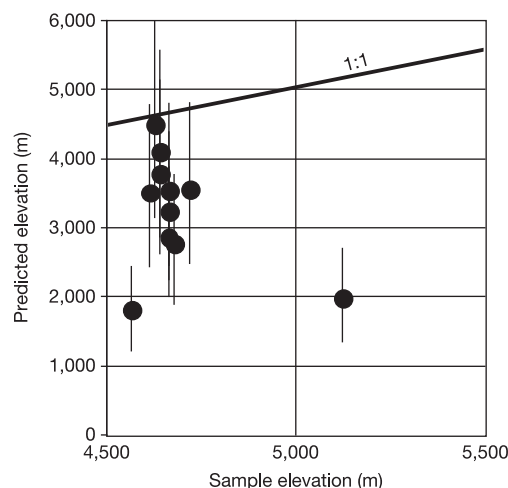


Figure 2 | Sample elevations versus predicted elevations of modern stream waters in central Tibet. Sample elevations are based on GPS estimates of elevations at the sample site. Predicted elevations (h) are based on measured oxygen isotopic composition differenced with the weighted mean annual composition of precipitation in New Delhi ($\delta^{18}\text{O}_p = -5.8\text{‰}$) and:

$$h = -6.14 \times 10^{-3} \Delta(\delta^{18}\text{O}_p)^4 - 0.6765 \Delta(\delta^{18}\text{O}_p)^3 - 28.623 \Delta(\delta^{18}\text{O}_p)^2 - 650.66 \Delta(\delta^{18}\text{O}_p) \quad (1)$$

The line of 1:1 correlation is included. Error bars are model-related 2σ uncertainties of elevation based on $\Delta(\delta^{18}\text{O}_w)$ (ref. 15).

importantly uncertainty in the temperature at which carbonate and water equilibrated, and the corresponding uncertainty in the estimate of the mean elevation; it also includes model-related uncertainties reflecting variations in temperature and relative humidity of low latitude and low elevation air masses¹⁵. Besides orographic ascent, probably the largest additional controller on $\Delta(\delta^{18}\text{O}_w)$ is evaporation. Evaporation preferentially removes ^{16}O , resulting in an enrichment of $\delta^{18}\text{O}_w$ in the remaining water and a reduction in $\Delta(\delta^{18}\text{O}_w)$. The best estimate of the Miocene palaeo-altitude of the Lunpola basin is thus provided by the most depleted $\Delta(\delta^{18}\text{O}_w)$ values, as these are probably the least affected by evaporative re-enrichment or other atmospheric processes that reduce the depletion relative to that predicted by the model. The most depleted samples with a $\Delta(\delta^{18}\text{O}_w)$ of $-10.0 \pm 2.0\text{‰}$ yield an estimated palaeo-elevation of 4,260 m $\pm$ 475/–575 m.

Upper Niubao Formation samples are derived from micritic CaCO_3 nodules from stacked argillic calcsols. Extensive sampling of multiple carbonate nodules from 12 different horizons at one locality (no. 305), as well as additional samples from overlying palaeosol carbonate nodule-bearing horizons, provide 60 separate analyses of $\delta^{18}\text{O}_c$ and $\delta^{13}\text{C}_c$. The average $\delta^{13}\text{C}_c$ of all of the samples is $-7.6 \pm 0.7\text{‰}$ (2σ), consistent with development associated with C_3 vegetation. $\delta^{18}\text{O}_c$ averages $-17.2 \pm 1.2\text{‰}$ (2σ). This consistency means that it does not matter whether the individual samples are grouped by locality or by horizon; as shown in Fig. 3, the results are similar. The mean $\Delta(\delta^{18}\text{O}_w)$ for these samples is $-12.6 \pm 1.2\text{‰}$ (2σ), corresponding with a predicted mean elevation of the basin of about 4,850 m $\pm$ 380/–460 m. Model uncertainty of elevation for a $\Delta(\delta^{18}\text{O}_w)$ of -12.6‰ is 4,850 m $\pm$ 1,630/–1,435 m (2σ) (ref. 15).

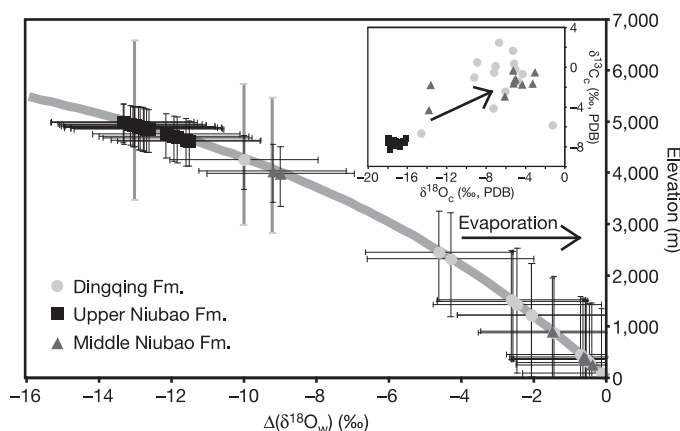


Figure 3 | $\Delta(\delta^{18}\text{O}_w)$ versus predicted elevations derived from Lunpola carbonates. Main panel, $\Delta(\delta^{18}\text{O}_w)$ versus predicted elevations of carbonate-derived estimates of local surface and soil waters from the Lunpola basin in central Tibet. Deriving $\delta^{18}\text{O}_w$ from $\delta^{18}\text{O}_c$ assumes that carbonate and water equilibrated at $10 \pm 10^\circ\text{C}$, corresponding to $\alpha_{\text{CaCO}_3-\text{H}_2\text{O}}$ of about $1.0323 + 0.0027/-0.0024$ for colder and warmer temperatures, respectively, based on the temperature dependence of the fractionation factor³², and we incorporate this uncertainty in our estimates of $\Delta(\delta^{18}\text{O}_w)$. The arrow emphasizes that evaporation reduces $\Delta(\delta^{18}\text{O}_w)$, which results in a lower estimate of predicted elevation. The model curve from Rowley *et al.*¹⁵ (based on equation (1)) is shown for reference. Two sets of uncertainties are plotted with each data point. Black, horizontal uncertainties combine uncertainties in $\delta^{18}\text{O}_c$, $\Delta(\delta^{18}\text{O}_w)$ and in the temperature of fractionation of carbonate and water. These uncertainties give rise to an uncertainty (1σ) in the estimate of the mean elevation of the sample, shown as the black vertical bars. Lighter grey vertical bars, shown only for three samples, reflect 2σ uncertainties associated with starting relative humidity and temperature of the low elevation moisture source (see ref. 15 for a more complete discussion). Inset, plot of $\delta^{13}\text{C}_c$ versus $\delta^{18}\text{O}_c$ of data from each of the sample horizons. The arrow shows the expected variation associated with evaporation. Symbols are the same for each plot.

Stratigraphically beneath the upper Niubao Formation, palaeosol carbonates are a sequence of variably dolomitized and silicified thin-bedded lacustrine carbonates of which we focused our sampling on limestones and marls. $\delta^{13}\text{C}_c$ ranges from -4.4‰ to -0.3‰ , while $\delta^{18}\text{O}_c$ ranges from -3.1‰ to -13.8‰ (PDB) (Fig. 3 inset). As with the Dingqing lacustrine carbonates, there is a substantial range in estimates of both $\Delta(\delta^{18}\text{O}_w)$ and predicted palaeo-elevations from these carbonates. The most negative samples, with $\Delta(\delta^{18}\text{O}_w)$ of about $-9.2 \pm 2.0\text{‰}$, give the best estimate of the palaeo-elevation for this sequence of about 4,050 m $\pm$ 510/–620 m. Model uncertainties corresponding with a $\Delta(\delta^{18}\text{O}_w)$ of about -9.2‰ are 4,050 m $\pm$ 1,420/–1,220 m (ref. 15).

Oxygen isotopic data from the late Eocene–Oligocene Niubao Formation and the Miocene Dingqing Formation yield remarkably consistent results. The most negative, and hence least potentially evaporatively enriched, samples yield estimated palaeo-altitudes for the Lunpola basin in excess of 4,000 m, and all overlap within uncertainty with the basin's current mean elevation of about 4,600 m. The upper Niubao Formation palaeosol carbonate nodules yield the most consistent results, suggesting a highly elevated basin within which C_3 vegetation dominated the carbon isotopic budget. Both older and younger lacustrine-dominated sequences appear to be more variably affected by evaporation, but nonetheless the most depleted oxygen isotopic compositions also have the most negative carbon isotopic compositions, consistent with limited evaporative or secondary alteration that suggest comparably high elevations. The most reasonable interpretation is that this part of the Tibetan plateau was already elevated to its current altitude by the time of deposition of the upper Niubao in late Eocene to early Oligocene times. Central to the above interpretations is the assumption that the stable-isotopic composition of Lunpola basin lacustrine and palaeosol carbonates accurately record the composition of meteoric waters at the time of deposition. Diagenetic alteration associated with high-temperature recrystallization or younger more depleted meteoric waters may result in a reduction in $\delta^{18}\text{O}$ compositions that may lead to erroneous interpretations of palaeo-elevation²³.

Several lines of evidence lead us to believe that overall diagenetic alteration of our samples has been minimal. First, all of our samples were derived from micritic carbonates that display little evidence of recrystallization. While the effects of diagenesis on carbonates can be subtle, our samples do not display the pervasive microcrystallization reported from the mid-Cenozoic basins of eastern Tibet²³. In addition, as a whole, our samples display a positive covariance of $\delta^{18}\text{O}$ and $\delta^{13}\text{C}$ values (Fig. 3 inset). We interpret this trend to reflect evaporative enrichment of meteoric waters in Lunpola lacustrine systems throughout deposition. Similar trends from other isotopic studies have been interpreted to represent lacustrine carbonates that have undergone minimal diagenetic alteration²³. The variability in isotopic composition does not display any vertical stratigraphic trends that might be attributed to burial diagenesis. In fact, the greatest observed $\delta^{18}\text{O}$ variations in the both the middle Niubao and Dingqing formations come from samples that are in close stratigraphic proximity (15 m and 50 cm, respectively). As such, it is unlikely that any diagenetic or more recent secondary alteration has effected the isotopic compositions of these carbonates.

Discussion

The palaeo-elevation estimates derived from the Lunpola basin are important for several reasons. First, these are currently the oldest data indicating that parts of the plateau were highly elevated ($>4\text{ km}$) by the late Eocene to early Oligocene. The consistency of the data from this basin over time supports models in which crustal rheology limits elevation and that once elevated the plateau surface remains at essentially the same elevation over time^{4,9,24}. Second, these data are consistent with the conclusion^{18,25} that models invoking significant mantle involvement through initial mantle-lithosphere thickening and subsequent convective destabilization and removal of this mantle

lithospheric root^{4–6} are not supported by any palaeo-elevation data available from the plateau.

An important question remains: are the estimates of the palaeo-elevation history of the Lunpola basin expected on the basis of current understanding of the controls on orogenic plateau development? This can only be addressed if explicit retrodictions of the growth of the plateau can be formulated. Figure 4 presents an estimate of the retrodicted increase in width of the plateau as a function of time. Our estimate derives from the model of Beaumont *et al.*³, as it makes the most explicit retrodictions currently available. The modelled³ width of the entire orogen in excess of 4 km elevation increases by about 40% of the cumulative convergence for the interval from ~54 Myr to 30 Myr after initiation of collision; this phase of the model includes erosion at the front, and a crust–mantle density contrast of 500 kg m⁻³. The entire width of the orogen is not preserved as a function of time owing to erosion, primarily along the Himalaya. We therefore measure width relative to the Indus–Yarlung Tsangpo suture, as it is more readily defined as a function of time and its position is tracked in the model³. Approximately 80% of the crustal mass of the plateau lies north of the suture and hence we further reduce the scaling of retrodicted plateau width by this factor, resulting in an overall scaling of plateau width north of the suture to 32% of the cumulative convergence. We use the independently determined India–Asia convergence history since the initiation of collision to scale the predicted width of the plateau in Fig. 4. Initiation of the India–Asia collision occurred at 50.6 ± 0.2 Myr ago²⁶ in the vicinity of Mount Everest. No data tightly constrain the age of initiation of collision farther east²⁷. Cumulative collision-related convergence varies as a function of position along the suture, but at 90° E the total is about 3,500 km (Fig. 4); it increases to about 3,700 km in the east and decreases to about 2,650 km in the west. Scaling by 40% predicts a current plateau width of 1,400 km and a width north of the suture of about 1,100 km, both of which accord well with observation. Between the time of initiation of the collision and magnetic reversal chron 21 (at ~46.2 Myr ago; ref. 28), the India–Asia convergence rate was approximately twice the subsequent rate, and hence the scaling requires an early, more rapid increase in width of the plateau, followed by a more steady northward growth at an average of about 21 km Myr⁻¹ (Fig. 4) since about 46.2 Myr ago.

The ages at which various basins situated north of the suture are retrodicted to have exceeded 4 km elevation are also plotted in Fig. 4.

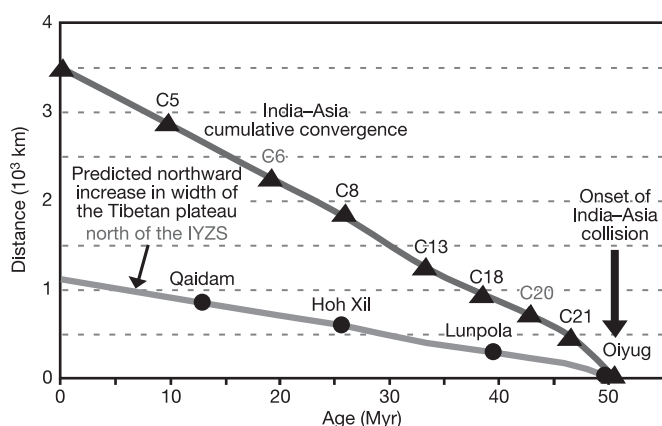


Figure 4 | Predicted elevation history of the Tibetan plateau north of the Indus–Yarlung Tsangpo suture (IYTS) at the longitude of 90° E, and the India–Asia convergence. Plateau width is scaled on the basis of thermo-mechanical modelling to the convergence history of the India–Asia collision. India–Asia convergence history is constrained at each of the magnetic reversals (filled triangles) represented by C5 through to C22 (ref. 28). Predicted ages of uplift of various basins above 4 km height across the plateau are shown by the filled circles. Locations of the basins are shown in Fig. 1.

Palaeo-elevation estimates are available for each of these basins, but unfortunately only the data from the Lunpola and Hoh Xil basins are derived from sufficiently old sediments to provide much of a test of the retrodictions. The Lunpola basin is retrodicted to have achieved an elevation in excess of 4 km by ~39 Myr ago, and the middle and upper Niubao data that date from the late Eocene to Oligocene (~35 ± 5 Myr ago) are compatible with this estimate (Fig. 4). Cyr *et al.*²⁹ report data from lacustrine limestones of the Fenghuoshan Group²⁹ of late Eocene age (~39 Myr ago) in the Hoh Xil basin (Fig. 1). Ca/Mg and C isotope data are consistent with these lakes being open and not significantly affected by evaporation, suggesting that their oxygen isotopic compositions should be reflective of in-flowing surface waters. Palaeo-elevation estimates from Fenghuoshan lacustrine carbonates suggest that this part of the plateau was at elevations below 2 km at 39 Myr ago, again consistent with the model which retrodicts elevation above 4 km of the Hoh Xil basin region at about 25 Myr ago. Unfortunately, there are as yet no younger materials that have yielded palaeo-elevation estimates available from this part of the plateau. The Qaidam basin is retrodicted to have been elevated above 4 km at about 13 Myr ago. The Qaidam basin is currently at an average elevation below 3 km, but shortening jumped across the Qaidam basin to the Qilian Shan at about this time⁸, isolating the Qaidam basin within the broader Tibetan orogenic system. Isotopic data from the Tarim and Qaidam basins are broadly consistent with changing orography and moisture source input at this time³⁰. A palaeo-enthalpy-based estimate²⁵ and an oxygen-isotope-based estimate¹⁸ agree with each other, and place the Oiyug basin (situated just north of the suture) above 4 km elevation at 15 Myr ago. The model retrodicts the uplift of this region shortly after collision and hence these data, although consistent with expectation, do not help to constrain the model.

Currently available data are consistent with a simple model that scales India–Asia convergence to increase in plateau width as a function of time. It is extremely encouraging that palaeo-elevation data from the Himalaya and Tibet are beginning to be able to discriminate among various models that have been proposed in the past. A significant finding in this regard is the absence of evidence for mantle lithosphere thickening and convective removal as a contributor to the elevation history of the plateau. At present, the predictions based on thermal-mechanical modelling³ are consistent with the existing data. Given our current data and understanding, estimates of plateau growth should be based on progressive northward growth of the plateau at between 4 km and 5 km elevation at a rate of about 20 km Myr⁻¹ after about 46.2 Myr ago. Regions farther to the east involve additional contributors to the strain accommodation history that make scaling of convergence to plateau width less straightforward. Regions that lie east of the convergence trajectory of India relative to Asia³¹ would not be expected to be elevated according to the simple model presented here. Eastward export of crustal mass, primarily through sub-crustal flow⁹, contributes a sink that does not exist in the centre of the plateau in the vicinity of Lunpola. Additional contributions from crustal and lithospheric escape may also be significant. Together, these suggest that increase in width in these more eastern regions may not be so simple. A final comment is to emphasize that there is no evidence supporting plateau-wide uplift at 8–10 Myr ago and hence any suggested correlates associated with previous model-derived⁶ suggestions now need to be re-examined.

Overview

Oxygen-isotope-based estimates of palaeo-altitude from the late Eocene–Oligocene Niubao Formation and Miocene middle Dingqing Formation of central Tibet indicate that the central Tibetan plateau has been characterized by elevations in excess of 4 km since 35 ± 5 Myr ago. Elevations in excess of 4 km are compatible with expectations if recent thermo-mechanical modelling³ are combined with estimates of the India–Asia age of collision and convergence to

provide first-order retrodictions of the elevation history of the Tibetan plateau. In the vicinity of the Lunpola basin, these retrodictions would imply that an elevated plateau existed by at least 39 Myr ago, and the data reported here are compatible with this. Data from the Lunpola basin further imply, again consistent with modelling, that once elevated the plateau's mean elevation did not vary within uncertainty. Finally, data from the Lunpola basin are consistent with observations derived from the Oiyug basin to the south^{18,25} that there is no evidence that the plateau underwent a plateau-wide uplift in the late Miocene (~8–10 Myr ago) associated with convective destabilization of a thickened lithospheric-mantle root beneath the plateau. Rather, all existing data are consistent with models in which the mantle lithosphere is simply subducted into the mantle.

Received 28 August; accepted 29 November 2005.

- Argand, E. La tectonique de l'Asie. In *Proc. 13th Int. Geological Congress* Vol. 1, Part 5, 171–372 (Vaillant-Carmanne, Liège, 1924).
- Dewey, J. F. & Burke, K. C. A. Tibetan, Variscan, and Precambrian basement reactivation: products of continental collision. *J. Geol.* **81**, 683–692 (1973).
- Beaumont, C., Jamieson, R. A., Nguyen, M. H. & Medvedev, S. Crustal channel flows: 1. Numerical models with applications to the tectonics of the Himalayan–Tibetan orogen. *J. Geophys. Res.* **109**, B06406, doi:10.1029/2003JB002809 (2004).
- England, P. & Houseman, G. Finite strain calculations of continental deformation. II: Application to the India–Asia plate collision. *J. Geophys. Res.* **91**, 3664–3676 (1986).
- Houseman, G. & England, P. in *The Tectonic Evolution of Asia* (eds Yin, A. & Harrison, T. M.) 3–17 (Cambridge Univ. Press, Cambridge, 1996).
- Molnar, P., England, P. & Martinod, J. Mantle dynamics, uplift of the Tibetan plateau, and the Indian Monsoon. *Rev. Geophys.* **31**, 357–396 (1993).
- Tapponnier, P., Peltzer, G., LeDain, A. Y., Armijo, R. & Cobbold, P. Propagating extrusion tectonics in Asia: new insights from simple experiments with plasticine. *Geology* **10**, 611–616 (1982).
- Tapponnier, P. et al. Oblique stepwise rise and growth of the Tibet plateau. *Science* **294**, 1671–1677 (2001).
- Royden, L. H. et al. Surface deformation and lower crustal flow in eastern Tibet. *Science* **276**, 788–790 (1997).
- Bureau of Geology and Mineral Resources Xizang Autonomous Region. *Regional Geology of Xizang (Tibet) Autonomous Region* (Geological Publishing House, Beijing, 1993).
- Zhai, G. et al. (eds) Oil and gas fields in Qinghai and Xizang. In *Petroleum Geology of China* Vol. 14 (Petroleum Industry Press, Beijing, 1990).
- Xia, J. B. Cenozoic of Baingoin and its borders, Xizang (Tibet) [Chinese with English summary]. *Contrib. Geol. Qinghai-Xizang (Tibet) Plateau* **3**, 243–254 (1983).
- Xu, Z. Y. The Tertiary and its petroleum potential in the Lunpola Basin, Tibet. *Oil Gas Geol.* **1**, 153–158 (1980).
- Xu, Z. Y., Zhao, J. P. & Wu, Z. L. On the Tertiary continental basins and their petroleum potential in Qinghai-Xizang (Tibet) Plateau with Lunpola Basin as example [Chinese with English summary]. *Contrib. Geol. Qinghai-Xizang (Tibet) Plateau* **17**, 391–399 (1985).
- Rowley, D. B., Pierrehumbert, R. T. & Currie, B. S. A new approach to stable isotope-based paleoaltimetry: Implications for paleoaltimetry and paleohypsometry of the High Himalaya since the Late Miocene. *Earth Planet. Sci. Lett.* **188**, 253–268 (2001).
- Garzzone, C. N., Dettman, D. L., Quade, J., DeCelles, P. G. & Butler, R. F. High times on the Tibetan Plateau: Paleoelevation of the Thakkhola graben, Nepal. *Geology* **28**, 339–342 (2000).
- Garzzone, C. N., Quade, J., DeCelles, P. G. & English, N. B. Predicting paleoelevation of Tibet and the Himalaya from $\delta^{18}\text{O}$ vs. altitude gradients in meteoric water across the Nepal Himalaya. *Earth Planet. Sci. Lett.* **183**, 215–229 (2000).
- Currie, B. S., Rowley, D. B. & Tabor, N. J. Middle Miocene paleoaltimetry of southern Tibet: Implications for the role of mantle thickening and delamination in the Himalayan orogen. *Geology* **33**, 181–184 (2005).
- Dansgaard, W. Stable isotopes in precipitation. *Tellus* **16**, 436–468 (1964).
- Rozanski, K., Araguas-Araguas, L. & Gonfiant, R. in *Climate Change in Continental Isotopic Records* 1–36 (American Geophysical Union, Washington, 1993).
- Tian, L., Masson-Delmotte, V., Stievenard, M., Yao, T. & Jouzel, J. Tibetan Plateau summer monsoon northward extent revealed by measurements of water stable isotopes. *J. Geophys. Res.* **106**, 28081–28088 (2001).
- IAEA. (<http://isohis.iaea.org/userupdate/GNIP2001yearly.xls>) (2001).
- Garzzone, C. N., Dettman, D. L. & Horton, B. K. Carbonate oxygen isotope paleoaltimetry: Evaluating the effect of diagenesis on paleoelevation estimates for the Tibetan plateau. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **212**, 119–140 (2004).
- Beaumont, C., Fullsack, P. & Hamilton, J. in *Thrust Tectonics* (ed. McClay, K. R.) 1–18 (Chapman & Hall, London, 1992).
- Spicer, R. A. et al. Constant elevation of southern Tibet over the past 15 million years. *Nature* **421**, 622–624 (2003).
- Zhu, B., Kidd, W. S. F., Rowley, D. B., Currie, B. S. & Shafique, N. Age of initiation of the India–Asia collision in the East-Central Himalaya. *J. Geol.* **113**, 265–285 (2005).
- Rowley, D. B. Age of initiation of collision between India and Asia: A review of stratigraphic data. *Earth Planet. Sci. Lett.* **145**, 1–13 (1996).
- Berggren, W. A., Kent, D. V., Swisher, C. C. III & Aubry, M. P. in *Time Scales and Global Stratigraphic Correlation* (eds Berggren, W. A., Kent, D. V., Aubry, M. P. & Hardenbohl, J.) 129–218 (Society of Economic Paleontologists and Mineralogists, Tulsa, Oklahoma, 1995).
- Cyr, A., Currie, B. S. & Rowley, D. B. Geochemical and stable isotopic evaluation of Fenghuoshan Group lacustrine carbonates, north-central Tibet: Implications for the paleoaltimetry of Late Eocene Tibetan Plateau. *J. Geol.* **113**, 517–533 (2005).
- Graham, S. A. et al. Stable isotope records of Cenozoic climate and topography, Tibetan Plateau and Tarim Basin. *Am. J. Sci.* **305**, 101–118 (2005).
- LePichon, X., Fournier, M. & Jolivet, L. Kinematics, topography, shortening and extrusion in the India–Eurasia collision. *Tectonics* **11**, 1085–1098 (1992).
- Friedman, I. & O'Neil, J. R. in *Data of Geochemistry* (ed. Fleischer, M.) 1–12 (US Geological Survey Prof. Pap. 440-K, Washington DC, 1977).

Acknowledgements Particular thanks are extended to D. Schrag for use of his laboratory at Harvard University for the oxygen and carbon isotopic analyses. We thank M. Haizhou of the Salt Lake Institute, Chinese Academy of Sciences, Xining, for logistical assistance and arranging access to sites in northern and central Tibet. We also thank A. Cyr for assistance in the field. D.B.R. thanks fellow members of the Earth System Evolution Program of the Canadian Institute for Advanced Research for discussion and comments. This work was funded by the National Science Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.B.R. (drowley@uchicago.edu).

ARTICLES

Restoring function in exhausted CD8 T cells during chronic viral infection

Daniel L. Barber¹, E. John Wherry², David Masopust¹, Baogong Zhu³, James P. Allison⁴, Arlene H. Sharpe⁵, Gordon J. Freeman³ & Rafi Ahmed¹

Functional impairment of antigen-specific T cells is a defining characteristic of many chronic infections, but the underlying mechanisms of T-cell dysfunction are not well understood. To address this question, we analysed genes expressed in functionally impaired virus-specific CD8 T cells present in mice chronically infected with lymphocytic choriomeningitis virus (LCMV), and compared these with the gene profile of functional memory CD8 T cells. Here we report that PD-1 (programmed death 1; also known as Pdc1) was selectively upregulated by the exhausted T cells, and that *in vivo* administration of antibodies that blocked the interaction of this inhibitory receptor with its ligand, PD-L1 (also known as B7-H1), enhanced T-cell responses. Notably, we found that even in persistently infected mice that were lacking CD4 T-cell help, blockade of the PD-1/PD-L1 inhibitory pathway had a beneficial effect on the 'helpless' CD8 T cells, restoring their ability to undergo proliferation, secrete cytokines, kill infected cells and decrease viral load. Blockade of the CTLA-4 (cytotoxic T-lymphocyte-associated protein 4) inhibitory pathway had no effect on either T-cell function or viral control. These studies identify a specific mechanism of T-cell exhaustion and define a potentially effective immunological strategy for the treatment of chronic viral infections.

Memory CD8 T-cell differentiation proceeds along distinct pathways after an acute versus a chronic viral infection^{1–3}. Memory CD8 T cells generated after an acute viral infection are highly functional and constitute an important component of protective immunity. In contrast, chronic infections are often characterized by varying degrees of functional impairment of virus-specific T-cell responses, and this defect is a principal reason for the inability of the host to eliminate the persisting pathogen. Although functional effector T cells are initially generated during the early stages of infection, they gradually lose function during the course of the chronic infection. This exhaustion of virus-specific T cells was first shown during persistent LCMV infection of mice^{4,5}. However, these findings were quickly extended to other model systems, as well as to chronic infections in humans, in particular human immunodeficiency virus (HIV), hepatitis B virus (HBV) and hepatitis C virus (HCV) infections—three chronic infections afflicting >500 million people worldwide^{6–8}.

PD-1 is upregulated during chronic viral infection

To examine the mechanism of T-cell dysfunction during chronic infection, we used the mouse model of infection with LCMV. An advantage of this model is the availability of LCMV strains that can cause either acute or chronic infections in adult mice; the Armstrong strain is cleared within a week, whereas the clone 13 strain establishes a persistent infection³. As these two strains differ in only two amino acids in the entire genome⁹, and neither of these mutations affects any of the known T-cell epitopes, it is possible to track the same CD8 T cell responses after an acute or chronic viral infection. In contrast to the highly robust memory CD8 T cells generated after an acute Armstrong infection, LCMV-specific CD8 T cells become exhausted during a persistent clone 13 infection^{3–5}. An example of impaired

cytokine production and proliferation by the exhausted CD8 T cells is shown in Fig. 1a. As a first step towards understanding the mechanisms of T-cell dysfunction, we performed a comparative genome-wide microarray analysis (Affymetrix) of genes expressed by exhausted versus functional LCMV-specific CD8 T cells of the same antigenic specificity. The most notable finding was the pronounced upregulation of messenger RNA encoding PD-1, an inhibitory receptor of the CD28 family^{10–13}, by the exhausted LCMV-specific CD8 T cells (Fig. 1b). This was confirmed at the protein level using antibodies specific for PD-1 (Fig. 1b). We found that high expression of this inhibitory receptor was a signature of the functionally exhausted CD8 T cells; all LCMV-specific CD8 T cells (comprising multiple epitopes) that were present in chronically infected mice expressed high levels of PD-1, whereas functional LCMV-specific memory CD8 T cells of the same antigenic specificities present in mice that had cleared the acute LCMV infection did not express any detectable levels of PD-1 (Fig. 1b and data not shown).

We next analysed the kinetics of PD-1 expression during both acute and chronic LCMV infection, and found that PD-1 was transiently expressed on early effector CD8 T cells after LCMV Armstrong infection but then was rapidly downregulated (Fig. 1c). In contrast, PD-1 expression continued to increase on virus-specific CD8 T cells in chronically infected mice and the high level of expression was sustained. PD-1 has two ligands—PD-L1 (B7-H1)^{14,15}, which is widely expressed on both haematopoietic and parenchymal cells, and PD-L2 (B7-DC)^{12,13}, which is predominantly expressed on macrophages and dendritic cells. We found that PD-L1 was expressed at very high levels in splenocytes from persistently infected mice, especially on virally infected cells (Fig. 1c). Thus, not only did the exhausted CD8 T cells express high levels of PD-1, but its

¹Emory Vaccine Center and Department of Microbiology and Immunology, Emory University School of Medicine, Atlanta, Georgia 30322, USA. ²Immunology Program, The Wistar Institute, Philadelphia, Pennsylvania 19104, USA. ³Department of Medical Oncology, Dana-Farber Cancer Institute, Department of Medicine, Harvard Medical School, Boston, Massachusetts 02115, USA. ⁴Department of Medicine, Howard Hughes Medical Institute, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA. ⁵Department of Pathology, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts 02115, USA.

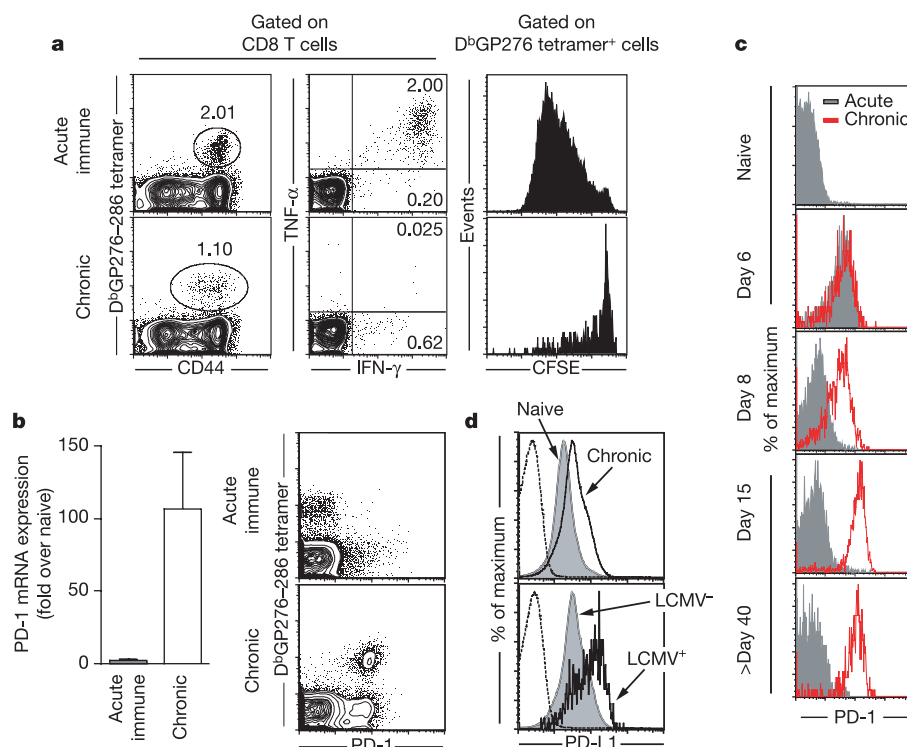


Figure 1 | PD-1 is highly upregulated by exhausted CD8 T cells during chronic infection. **a**, GP276-specific CD8 T cells from LCMV Armstrong immune and clone 13 chronically infected mice (>60 days post-infection) were stimulated with peptide and analysed for cytokine production (5 h) and proliferation (60 h). **b**, PD-1 mRNA (Affymetrix array) and protein expression in virus-specific CD8 T cells isolated from acutely or chronically infected mice (>60 days post-infection). **c**, Kinetics of PD-1 expression on

D^bGP276–286 tetramer-positive CD8 T cells during acute (Armstrong) and chronic (clone 13) infection. The top plot is gated on CD8 T cells from a naive mouse. **d**, PD-L1 surface expression on splenocytes from naive and chronically infected mice (top plot) and on LCMV-infected or uninfected splenocytes from a chronically infected mouse (bottom plot). Error bars are s.e.m. throughout. Numbers on FACS plots indicate the percentage of cells in each region.

ligand was also upregulated on infected cells, suggesting a role for this inhibitory pathway in regulating T-cell function during chronic LCMV infection.

PD-L1 blockade enhances viral control and T-cell responses

To test the above hypothesis, we treated persistently infected mice with anti-PD-L1 blocking antibody and monitored T-cell responses

and viral control (Fig. 2). We found that virus-specific CD8 T-cell responses were enhanced both quantitatively and qualitatively by the anti-PD-L1 blockade. The treated mice showed significant increases ($P = 0.02$) in the numbers of LCMV-specific CD8 T cells, as measured by three different MHC class I tetramers (Fig. 2a). Most notably, the T cells became functionally superior following anti-PD-L1 treatment, and a higher fraction of the virus-specific CD8 T cells were capable of

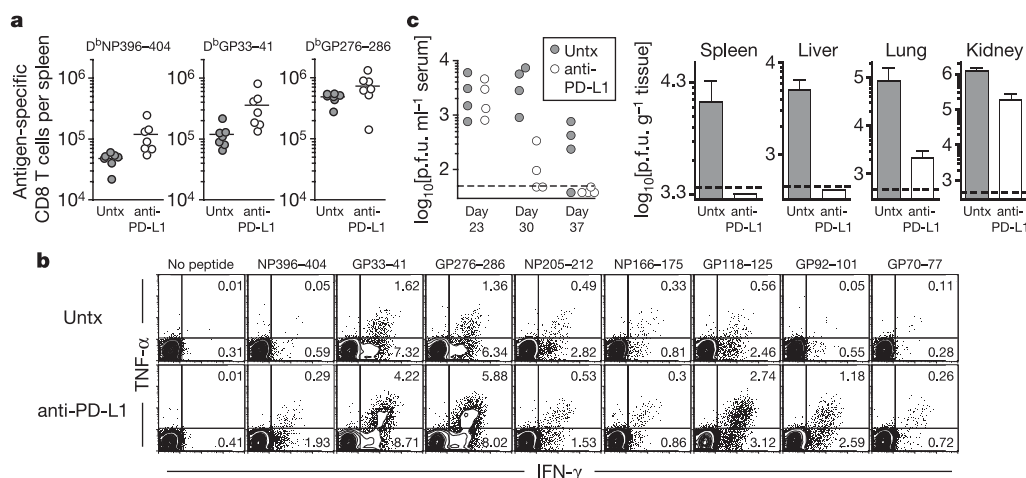


Figure 2 | Blockade of the PD-1/PD-L1 inhibitory pathway increases CD8 T-cell function and enhances the clearance of chronic viral infection. LCMV clone-13-infected mice were treated with anti-PD-L1 from day 23 to day 37 post-infection. **a**, Virus-specific CD8 T-cell numbers in the spleen. Results are pooled from two experiments (NP396 and GP33; $P = 0.02$).

b, Cytokine production by splenocytes after stimulation with the indicated peptides. **c**, Viral titres in the serum ($P = 0.0016$ at day 30 and $P = 0.035$ at day 37) and indicated tissues at day 37 (spleen $P = 0.003$, liver $P = 0.0084$, lung $P = 0.0076$, kidney $P < 0.0001$). $n = 4$ mice per group. Data are representative of two experiments. Untx, untreated.

producing both interferon (IFN)- γ and tumour necrosis factor (TNF)- α compared to untreated mice ($P = 0.01$) (Fig. 2b). This improved functionality was seen with CD8 T cells responding to multiple epitopes (Fig. 2b). Significantly, PD-L1 blockade resulted in a substantial reduction in virus levels; the treated mice had cleared infectious virus from the serum, spleen and liver, while untreated mice still harboured high levels of virus in these tissues ($P = 0.01$) (Fig. 2c). Also, the anti-PD-L1-treated mice that resolved the infection remained healthy, and did not exhibit any overt signs of disease.

As PD-1 was transiently upregulated during acute LCMV infection, we treated Armstrong-infected mice with anti-PD-L1 antibody during the first week of infection and monitored the T-cell response. We found that PD-L1 blockade did not increase virus-specific CD8 T-cell responses in the acute infection model (data not shown). Thus, it appears that the PD-1 inhibitory pathway is particularly important in exhausted T cells.

Restoring function in 'helpless' CD8 T cells

We next asked whether blockade of the PD-1/PD-L1 inhibitory pathway would enhance CD8 T-cell responses in the absence of CD4 T cell help. This is a much more stringent model of chronic infection and the 'helpless' CD8 T cells show even more pronounced functional defects^{4,16,17}. Even under these conditions, when CD4 T cell help was lacking, PD-L1 blockade was highly effective in increasing T-cell numbers, improving their quality, and reducing viral load (Fig. 3). There were approximately sevenfold more D^bGP276–286-specific CD8 T cells ($P < 0.0001$), and fourfold more D^bGP33–41-specific cells ($P < 0.0001$) in the spleens of anti-PD-L1-treated mice than untreated control mice (Fig. 3a, b). There was also a substantial increase in virus-specific CD8 T cells in non-lymphoid tissues such as liver and lung and even in intraepithelial lymphocytes (IEL) (Fig. 3c). To determine whether this increase in cell numbers was due to cell division, we measured 5-bromodeoxyuridine (BrdU) incorporation during the period of anti-PD-L1 treatment (Fig. 3d). In untreated mice, the majority (>55%) of D^bGP276–286-specific CD8 T cells did not incorporate any BrdU, while in treated mice, $\geq 90\%$ of these cells were positive with much higher levels (mean fluorescence intensity) of BrdU incorporation. In addition to the increased number of BrdU-positive cells, we also found an increased number of cells expressing Ki67, a protein associated with cell-cycle progression (19% in untreated versus 60% in anti-PD-L1 treated mice) (Fig. 3d).

We next measured the effect of anti-PD-L1 blockade on the ability of virus-specific CD8 T cells to produce IFN- γ in the absence of CD4 T cell help. There was a substantial increase in the number of IFN- γ -producing cells in the treated mice (Supplementary Fig. 1). To compare the function of CD8 T cells in untreated and anti-PD-L1-treated mice on a per-cell basis, we calculated the percentage of D^bGP276–286-tetramer-positive cells that were capable of producing IFN- γ . In untreated mice, only $\sim 20\%$ of D^bGP276–286-specific CD8 T cells were capable of producing IFN- γ , whereas in anti-PD-L1 treated mice, $\geq 50\%$ of these CD8 T cells produced IFN- γ ($P < 0.0001$) (Fig. 3e).

During chronic LCMV infection, exhausted virus-specific cytotoxic T lymphocytes (CTLs) are not able to lyse targets in ⁵¹Cr-release assays (Fig. 3f and refs 3–5). Direct *ex vivo* lytic ability became detectable, however, after anti-PD-L1 blockade (Fig. 3f). We also measured the ability of virus-specific CD8 T cells to degranulate by monitoring the appearance of lysosomal markers CD107a/b on the cell surface after re-stimulation with peptide. There was a dramatic increase in the number of virus-specific cells capable of degranulating in anti-PD-L1-treated mice (Fig. 3g). To test whether the increase in IFN- γ production and lytic activity corresponded to better viral control, we measured viral titres in the spleen, liver, lung and serum. As shown in Fig. 3h, there were significant reductions in virus levels in the spleen ($P = 0.008$), liver ($P < 0.0001$), lung ($P = 0.0002$) and serum ($P = 0.003$) in the treated mice. Thus, the results shown in

Fig. 3 demonstrate that PD-L1 blockade can restore function even in 'helpless' exhausted CD8 T cells.

To more directly assess the role of PD-1 in T-cell dysfunction, we treated chronically infected mice with a blocking antibody against PD-1 itself. We found that PD-1 blockade also restored function in the exhausted CD8 T cells (Supplementary Fig. 2). However, it is worth noting that blockade with anti-PD-L1 antibody was more effective than anti-PD-1 antibody in enhancing T-cell responses in the chronically infected mice. Potential explanations for this finding include blocking efficiency, differential expression of PD-L1 and PD-L2, or the possibility that PD-L2 might provide 'positive' co-stimulatory signals^{18,19}.

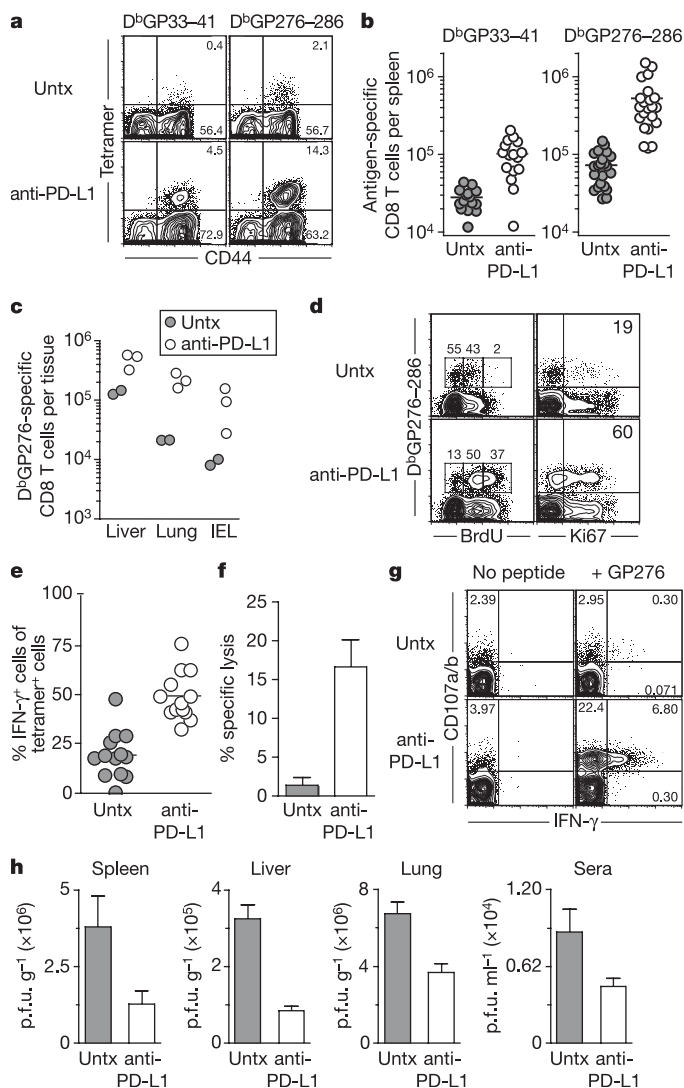


Figure 3 | Restoration of function in 'helpless' exhausted CD8 T cells. Mice were depleted of CD4 T cells and then infected with LCMV clone 13. After at least 40 days, chronically infected mice were treated with anti-PD-L1 for two weeks. Frequency and total number of tetramer-positive CD8 T cells in the spleen (a, b) or indicated tissues (c) ($P < 0.0001$ for GP33 and GP276). d, BrdU incorporation and Ki67 expression. $n > 7$ mice per group. e, Percentage of D^bGP276–286 tetramer-positive cells capable of producing IFN- γ ($P < 0.0001$). f, Direct *ex vivo* ⁵¹Cr-release activity using target cells coated with GP33–41, GP276–286 and GP118–125 peptides (effector:target ratio of 100:1). g, To measure degranulation by LCMV-specific CD8 T cells, splenocytes were stimulated with peptide in the presence of anti-CD107a/b antibodies and then co-stained for IFN- γ . h, Viral titres in the indicated tissues. $n > 10$ mice. (spleen $P = 0.008$, liver $P < 0.0001$, lung $P = 0.0002$, serum $P = 0.03$).

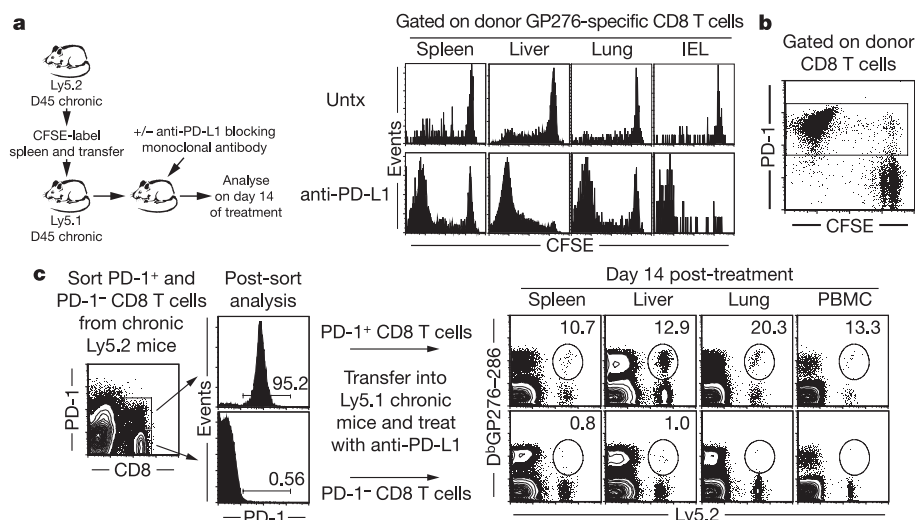


Figure 4 | Proliferation of exhausted CD8 T cells in both lymphoid and non-lymphoid tissues after anti-PD-L1 treatment. **a**, Ly5.1 and Ly5.2 mice were depleted of CD4 T cells and infected with LCMV clone 13. Splenocytes from chronically infected Ly5.2 mice were CFSE-labelled and transferred into chronically infected Ly5.1 mice. Recipient mice were then treated with anti-PD-L1 for two weeks, and proliferation of virus-specific donor CD8 T cells was analysed in the indicated tissues. **b**, PD-1 expression on

proliferating donor CD8 T cells from panel **a**. **c**, PD-1-positive and -negative CD8 T cells from Ly5.2 chronic mice were sorted and transferred into Ly5.1 chronic mice. Mice were then treated with anti-PD-L1 and killed after 14 days. Donor-derived tetramer-positive cells were measured in the indicated tissues. Plots are gated on CD8 T cells. Numbers represent the percentage of donor CD8 T cells specific for D^bGP276–286.

CTLA-4 is another inhibitory receptor of the CD28 family, and several studies have shown that treatment with anti-CTLA-4 blocking antibodies can enhance T-cell immunity against tumours²⁰. Interestingly, the exhausted LCMV-specific CD8 T cells also expressed elevated levels of CTLA-4 mRNA (data not shown). However, in contrast to our results with anti-PD-L1 or anti-PD-1 antibody blockade, treatment of chronically infected mice with anti-CTLA-4 blocking antibody had no effect on either virus-specific T-cell responses or viral control. Also, there were no synergistic effects of co-blockade with anti-PD-L1 and anti-CTLA-4 (Supplementary Fig. 3).

Proliferation of exhausted CD8 T cells

The expansion of exhausted CD8 T cells was one of the most remarkable effects after blockade of the PD-1/PD-L1 inhibitory pathway. To investigate this further, we used carboxyfluorescein diacetate, succinimidyl ester (CFSE)-labelled splenocytes from chronically infected Ly5.2 mice, transferred them into chronically infected Ly5.1 mice, and then treated the mice with anti-PD-L1 antibody (Fig. 4a). The use of this congenic system (Ly5.1 versus Ly5.2 mice chronically infected with LCMV) allowed us to distinguish between donor and host T cells and directly monitor the proliferation of the CFSE-labelled exhausted virus-specific CD8 T cells. When the exhausted cells (Ly5.2) were transferred back into a chronically infected mouse (Ly5.1), they underwent minimal to no proliferation in any of the tissues examined (spleen, liver, lung or IEL) (Fig. 4a). In contrast, in the presence of anti-PD-L1 blocking antibody, there was substantial proliferation and expansion of the exhausted virus-specific CD8 T cells in all of the tissues analysed (Fig. 4a). Furthermore, many of the donor cells recovered from both lymphoid and non-lymphoid tissues had diluted CFSE to background levels, showing they had divided more than eight times (Fig. 4a). It is worth noting that all exhausted virus-specific CD8 T cells that had undergone this rapid (antigen-driven) proliferation after anti-PD-L1 blockade still expressed high levels of the inhibitory receptor PD-1 (Fig. 4b). This suggests that continuous engagement of PD-L1 with PD-1 was the key event in inhibiting proliferation of the exhausted CD8 T cells. Once this interaction was disrupted by the anti-PD-L1 blocking antibody, the 'brake' was released and the exhausted CD8 T cells could proliferate in response to viral antigen.

Although our data so far make a compelling argument for PD-1/PD-L1 blockade restoring function in exhausted CD8 T cells, there is a possibility that the enhanced immune responses we are seeing in the chronically infected mice after anti-PD-L1 or anti-PD-1 treatment are actually due to the generation of *de novo* T-cell responses from 'naïve' CD8 T cells²¹. To address this issue, we sorted PD-1⁺ and PD-1⁻ CD8 T cells from chronically infected Ly5.2 mice, transferred them into chronically infected Ly5.1 mice, and then treated these mice with anti-PD-L1 antibody. We found that anti-PD-L1 treatment only enhanced immune responses of the sorted PD-1⁺ CD8 T cells (the cell population containing exhausted cells), and had no effect on PD-1⁻ CD8 T cells (containing naïve cells). These data show unequivocally that PD-L1 blockade restores function in the exhausted CD8 T cells (Fig. 4c).

It was of interest to determine whether the increased numbers of virus-specific CD8 T cells would be sustained after cessation of the anti-PD-L1 treatment. As shown in Supplementary Fig. 4, T-cell numbers remained stable and more functional for several weeks following the transient blockade. This shows that the expanded population of virus-specific CD8 T cells does not suddenly decline after the antibody treatment has been stopped. This is an important point, because it sets the stage for future studies designing intermittent blockade therapy for increasing, in a step-wise manner, the numbers of virus-specific T cells.

Infection of PD-L1^{-/-} mice

A prediction from our antibody blockade experiments is that T cells should not exhaust in PD-L1^{-/-} mice after infection with LCMV clone 13. To test this, we infected PD-L1^{-/-} mice with either Armstrong or clone 13. When PD-L1^{-/-} mice were infected with the LCMV Armstrong strain (acute infection), they behaved just like wild-type mice, producing normal CD8 T cell responses (Fig. 5a) and controlling the infection. In contrast, when PD-L1^{-/-} mice were infected with LCMV clone 13 they died owing to immunopathologic damage. This result tells us two important things: one, that the PD-1/PD-L1 axis operates primarily under conditions of sustained high levels of antigenic stimulation; and second, that the PD-1 inhibitory pathway may have evolved to regulate immune-mediated damage during a persistent infection by turning 'off' the virus-specific T cells. The pathogen ends up taking advantage of this inhibitory pathway to

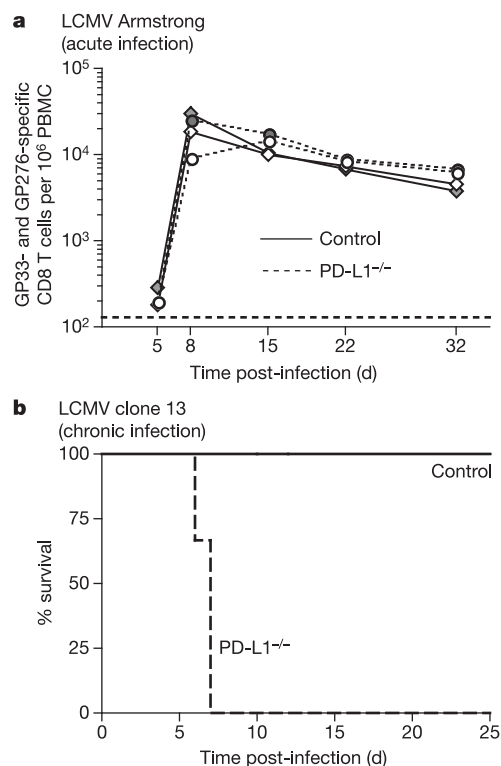


Figure 5 | PD-L1^{-/-} mice make normal CD8 T-cell responses after LCMV Armstrong infection but succumb to clone 13 infection owing to immunopathology. **a**, PD-L1^{-/-} and littermate control mice were infected with LCMV Armstrong, and virus-specific CD8 T cells were analysed in the peripheral blood mononuclear cells (PBMC) using the tetramers GP33 (diamonds) and GP276 (circles). **b**, PD-L1^{-/-} or littermate control mice were infected with clone 13 and survival was monitored.

establish persistence in the host. However, as we have shown, one can turn the tables on the pathogen by strategically blocking this inhibitory pathway and turning the T cells back on. However, one needs to carefully monitor the potential for autoimmunity and immunopathologic damage under such conditions.

Conclusion

The PD-1 inhibitory pathway is known to regulate immune responses to self antigens^{11–13}. Our study now defines a new and critical role for this pathway in modulating T-cell function during chronic viral infection. In addition, we show that PD-1 regulates the distinct pathways of memory CD8 T-cell differentiation observed during acute versus chronic viral infection. PD-1 expression was dramatically upregulated after the effector T-cell stage in chronically infected mice, and was rapidly downregulated in mice that cleared the infection (Fig. 1 and data not shown). It has been suggested that T cells adapt to persistent antigen by downregulating the responsiveness of their T-cell antigen receptor (TCR) signalling machinery^{22,23}, so defects in virus-specific CD8 T-cell function during chronic infection may be due to inhibition of TCR signal transduction. Inhibition mediated by PD-1 requires close proximity of PD-1 to the site of TCR engagement and does not signal in the absence of a TCR signal. Following crosslinking by PD-1 ligand, the immunoreceptor tyrosine-based switch motif (ITSM) in the cytoplasmic domain of PD-1 is phosphorylated and recruits the phosphatases SHP-1 and SHP-2. These phosphatases act on proximal signalling kinases of the TCR pathway, reducing the TCR signal and leading to diminished T-cell activation and cytokine production^{12,13}. Therefore, under conditions of persistent antigen, T cells may modulate their responsiveness by upregulating inhibitory receptors such as PD-1 that attenuate TCR signalling.

Although T-cell dysfunction is a common feature of many chronic viral infections, the underlying mechanisms have remained poorly understood. We have now not only identified a major inhibitory pathway operating during chronic LCMV infection, a classic model of viral persistence in its natural host, but have also developed a simple and highly effective strategy for restoring function in the exhausted CD8 T cells and enhancing viral control. It remains to be determined whether these findings can be extended to other models of chronic infections, and especially to persistent infections of humans. From this perspective, it is worth noting that a study by Iwai *et al.*²⁴ has shown that PD-1-knockout mice exhibit better control of adenovirus infection, and a recent study²⁵ using HBV transgenic mice suggests a role for PD-1 in regulating T-cell responses. Thus, it is likely that the immunological strategy we have developed for restoring function in exhausted T cells may have broader applications. Also, our finding that PD-1/PD-L1 blockade was effective in enhancing CD8 T-cell responses even under conditions of CD4 T-cell deficiency is of particular relevance to the situation seen during HIV infection^{6,7}.

Finally, our studies have implications for therapeutic vaccination and T-cell immunotherapy^{26,27}. Despite their enormous potential, therapeutic vaccines have had minimal to no success in eliminating chronic infections. One of the main reasons for their failure is the limited proliferative potential of the exhausted T cells²⁸. Our studies now provide a potential strategy for improving the efficacy of therapeutic vaccination by combining it with PD-1/PD-L1 blockade. Similarly, the efficacy of T-cell therapy for chronic infections or tumours may be enhanced by blocking the PD-1 inhibitory pathway.

METHODS

Mice and infections. Ly5.2 C57BL/6, Ly5.1 C57BL/6, 129 PD-L1^{-/-} and 129 littermate control mice were used in this study. LCMV Armstrong and clone 13 infections were performed as described³. Where indicated, CD4 T cells were depleted as described⁴.

In vivo antibody blockade. Two hundred micrograms of rat anti-mouse PD-L1 (10F5C5 or 10F9G2; ref. 29), rat IgG2b isotype control, rat anti-mouse PD-1 (29F.1A12) or anti-mouse CTLA-4 (UC10-4F10-11 or 9D9.1.1.7) were administered intraperitoneally every third day. The two different anti-PD-L1 clones gave indistinguishable results and the isotype control had no effect.

Gene array analysis. Naive P14 transgenic CD8 T cells, D^bGP33–41-specific memory CD8 T cells from Armstrong immune mice, and D^bGP33–41-specific or D^bGP276–286-specific CD8 T cells from CD4-depleted clone-13-infected mice were purified by fluorescence-activated cell sorting (FACS), and RNA isolation and gene array analysis was performed as previously described³⁰.

Intracellular cytokine, degranulation and ⁵¹Cr-release assays. Intracellular cytokine staining was performed as described³. To detect degranulation, splenocytes were stimulated for 5 h in the presence of brefeldin, monensin, anti-CD107a-FITC and anti-CD107b-FITC. ⁵¹Cr-release assays were performed as described³.

Received 1 August; accepted 21 November 2005.

Published online 28 December 2005.

- Klenerman, P. & Hill, A. T cells and viral persistence: lessons from diverse infections. *Nature Immunol.* **6**, 873–879 (2005).
- Wherry, E. J., Barber, D. L., Kaech, S. M., Blattman, J. N. & Ahmed, R. Antigen-independent memory CD8 T cells do not develop during chronic viral infection. *Proc. Natl Acad. Sci. USA* **101**, 16004–16009 (2004).
- Wherry, E. J., Blattman, J. N., Murali-Krishna, K., van der Most, R. & Ahmed, R. Viral persistence alters CD8 T-cell immunodominance and tissue distribution and results in distinct stages of functional impairment. *J. Virol.* **77**, 4911–4927 (2003).
- Zajac, A. J. *et al.* Viral immune evasion due to persistence of activated T cells without effector function. *J. Exp. Med.* **188**, 2205–2213 (1998).
- Gallimore, A. *et al.* Induction and exhaustion of lymphocytic choriomeningitis virus-specific cytotoxic T lymphocytes visualized using soluble tetrameric major histocompatibility complex class I-peptide complexes. *J. Exp. Med.* **187**, 1383–1393 (1998).
- Pantaleo, G. & Koup, R. A. Correlates of immune protection in HIV-1 infection: what we know, what we don't know, what we should know. *Nature Med.* **10**, 806–810 (2004).
- Letvin, N. L. & Walker, B. D. Immunopathogenesis and immunotherapy in AIDS virus infections. *Nature Med.* **9**, 861–866 (2003).

8. Rehmann, B. & Nascimbeni, M. Immunology of hepatitis B virus and hepatitis C virus infection. *Nature Rev. Immunol.* **5**, 215–229 (2005).
9. Matloubian, M., Kolhekar, S. R., Somasundaram, T. & Ahmed, R. Molecular determinants of macrophage tropism and viral persistence: importance of single amino acid changes in the polymerase and glycoprotein of lymphocytic choriomeningitis virus. *J. Virol.* **67**, 7340–7349 (1993).
10. Ishida, Y., Agata, Y., Shibahara, K. & Honjo, T. Induced expression of PD-1, a novel member of the immunoglobulin gene superfamily, upon programmed cell death. *EMBO J.* **11**, 3887–3895 (1992).
11. Nishimura, H., Nose, M., Hiai, H., Minato, N. & Honjo, T. Development of lupus-like autoimmune diseases by disruption of the PD-1 gene encoding an ITIM motif-carrying immunoreceptor. *Immunity* **11**, 141–151 (1999).
12. Sharpe, A. H. & Freeman, G. J. The B7–CD28 superfamily. *Nature Rev. Immunol.* **2**, 116–126 (2002).
13. Chen, L. Co-inhibitory molecules of the B7–CD28 family in the control of T-cell immunity. *Nature Rev. Immunol.* **4**, 336–347 (2004).
14. Freeman, G. J. *et al.* Engagement of the PD-1 immunoinhibitory receptor by a novel B7 family member leads to negative regulation of lymphocyte activation. *J. Exp. Med.* **192**, 1027–1034 (2000).
15. Dong, H., Zhu, G., Tamada, K. & Chen, L. B7-H1, a third member of the B7 family, co-stimulates T-cell proliferation and interleukin-10 secretion. *Nature Med.* **5**, 1365–1369 (1999).
16. Matloubian, M., Concepcion, R. J. & Ahmed, R. CD4⁺ T cells are required to sustain CD8⁺ cytotoxic T-cell responses during chronic viral infection. *J. Virol.* **68**, 8056–8063 (1994).
17. Bevan, M. J. Helping the CD8⁺ T-cell response. *Nature Rev. Immunol.* **4**, 595–602 (2004).
18. Shin, T. *et al.* *In vivo* costimulatory role of B7-DC in tuning T helper cell 1 and cytotoxic T lymphocyte responses. *J. Exp. Med.* **201**, 1531–1541 (2005).
19. Oflazoglu, E. *et al.* Paradoxical role of programmed death-1 ligand 2 in Th2 immune responses *in vitro* and in a mouse asthma model *in vivo*. *Eur. J. Immunol.* **34**, 3326–3336 (2004).
20. Egen, J. G., Kuhns, M. S. & Allison, J. P. CTLA-4: new insights into its biological function and use in tumour immunotherapy. *Nature Immunol.* **3**, 611–618 (2002).
21. Probst, H. C., McCoy, K., Okazaki, T., Honjo, T. & van den Broek, M. Resting dendritic cells induce peripheral CD8⁺ T cell tolerance through PD-1 and CTLA-4. *Nature Immunol.* **6**, 280–286 (2005).
22. Tanchot, C. *et al.* Modifications of CD8⁺ T cell function during *in vivo* memory or tolerance induction. *Immunity* **8**, 581–590 (1998).
23. Singh, N. J. & Schwartz, R. H. The strength of persistent antigenic stimulation modulates adaptive tolerance in peripheral CD4⁺ T cells. *J. Exp. Med.* **198**, 1107–1117 (2003).
24. Iwai, Y., Terawaki, S., Ikegawa, M., Okazaki, T. & Honjo, T. PD-1 inhibits antiviral immunity at the effector phase in the liver. *J. Exp. Med.* **198**, 39–50 (2003).
25. Isogawa, M., Furuichi, Y. & Chisari, F. V. Oscillating CD8⁺ T cell effector functions after antigen recognition in the liver. *Immunity* **23**, 53–63 (2005).
26. Dudley, M. E. & Rosenberg, S. A. Adoptive-cell-transfer therapy for the treatment of patients with cancer. *Nature Rev. Cancer* **3**, 666–675 (2003).
27. Autran, B., Carcelain, G., Combadiere, B. & Debre, P. Therapeutic vaccines for chronic infections. *Science* **305**, 205–208 (2004).
28. Wherry, E. J., Blattman, J. N. & Ahmed, R. Low CD8 T-cell proliferative potential and high viral load limit the effectiveness of therapeutic vaccination. *J. Virol.* **79**, 8960–8968 (2005).
29. Rodig, N. *et al.* Endothelial expression of PD-L1 and PD-L2 down-regulates CD8⁺ T cell activation and cytotoxicity. *Eur. J. Immunol.* **33**, 3117–3126 (2003).
30. Kaech, S. M., Hemby, S., Kersh, E. & Ahmed, R. Molecular and functional profiling of memory CD8 T cell differentiation. *Cell* **111**, 837–851 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Y. Blinder and M. Hulsey for technical assistance, and members of the Ahmed laboratory for helpful discussions. R.A., E.J.W., A.H.S. and G.J.F. were supported by NIH grants and the Gates Grand Challenges in Global Health, and J.P.A. by The Howard Hughes Medical Institute and NIH grants. E.J.W. and D.M. were supported by the Cancer Research Institute.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.A. (ra@microbio.emory.edu).

ARTICLES

Structural mechanism of plant aquaporin gating

Susanna Törnroth-Horsefield¹, Yi Wang², Kristina Hedfalk¹, Urban Johanson³, Maria Karlsson³, Emad Tajkhorshid², Richard Neutze¹ & Per Kjellbom³

Plants counteract fluctuations in water supply by regulating all aquaporins in the cell plasma membrane. Channel closure results either from the dephosphorylation of two conserved serine residues under conditions of drought stress, or from the protonation of a conserved histidine residue following a drop in cytoplasmic pH due to anoxia during flooding. Here we report the X-ray structure of the spinach plasma membrane aquaporin SoPIP2;1 in its closed conformation at 2.1 Å resolution and in its open conformation at 3.9 Å resolution, and molecular dynamics simulations of the initial events governing gating. In the closed conformation loop D caps the channel from the cytoplasm and thereby occludes the pore. In the open conformation loop D is displaced up to 16 Å and this movement opens a hydrophobic gate blocking the channel entrance from the cytoplasm. These results reveal a molecular gating mechanism which appears conserved throughout all plant plasma membrane aquaporins.

Water is the medium of life. Because biological membranes have only limited intrinsic water permeability, cells facilitate the flux of water into and out of the cell by means of a family of water-specific membrane protein channels called aquaporins¹. Members of the aquaporin family are found in archaea, eubacteria and eukaryotes, including fungi, animals and plants. They serve an astonishing variety of physiological functions^{2–4} and are easily identified by sequence similarity across all kingdoms of life. In higher eukaryotes, aquaporins are frequently regulated by phosphorylation, pH and osmolarity^{2–5}. Aquaporins in plants and animals are highly conserved and form large protein families with 35 members in higher plants⁶ and 13 members in humans^{3,4,7}. Phylogenetic analyses divide the plant aquaporins into four subfamilies (Supplementary Fig. S1) and their presence in primitive plants such as the bryophyte *Physcomitrella patens* implies that this specialization is ancient⁸. In the model plant *Arabidopsis thaliana* there are 13 well-conserved plasma membrane aquaporins (plasma membrane intrinsic proteins or PIPs) and these further separate into two distinct phylogenetic groups (PIP1 and PIP2).

In plants a large osmotic gradient between the cell interior and exterior can be maintained as a result of the strength of the cellulose microfibril cell wall. The resulting turgor pressure is crucial for many aspects of plant physiology and helps to give plant cells their structure and rigidity. Land plants have evolved to cope with rapid changes in the availability of water by regulating all aquaporins that lie within the plasma membrane (Fig. 1). Closure of the plant plasma membrane aquaporin SoPIP2;1 of spinach (formerly called PM28A (ref. 9)) is triggered by the dephosphorylation of two serine residues: Ser 115 in the cytosolic loop B (conserved as Ser in 12 of the 13 *Arabidopsis* PIPs, and as Thr in one of them) and Ser 274 in the carboxy-terminal region^{9,10} (fully conserved in all eight *Arabidopsis* PIP2s). Both residues are situated in consensus phosphorylation sites (Supplementary Fig. S2). Furthermore, the simultaneous closure of all *Arabidopsis* PIPs upon anoxia was recently reported to depend on the protonation of a strictly conserved histidine residue in loop D

(ref. 11), which corresponds to His 193 in SoPIP2;1. Distinct chemical signals acting on residues well separated in sequence induce an identical physiological response within PIPs.

A number of high-resolution structures have been reported for water^{12–16} and glycerol^{17,18} channels, and the mechanism of water and glycerol permeation has been extensively studied with molecular dynamics simulations^{18–21}. Nevertheless, plant aquaporin structures have been reported only at low resolution^{22,23}, and no gating mechanism has been unambiguously demonstrated. Furthermore, the correct assignment of the conformational state of the putative pH-gated channel aquaporin 0 (AQP0) has been controversial^{15,16}. To shed light on the molecular mechanism of aquaporin gating, we crystallized and solved the structure of SoPIP2;1 to 2.1 Å resolution, capturing it in its closed conformation. At this resolution the detailed interactions of all residues implicated in gating^{9–11} could

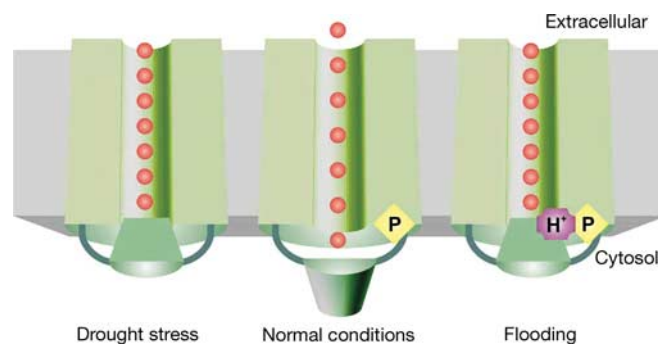


Figure 1 | Diagram illustrating the structural mechanism of aquaporin gating in plant plasma membranes. During drought stress the PIPs close in response to the dephosphorylation of two highly conserved serine residues (Ser 115 and Ser 274 of SoPIP2;1), whereas during flooding they close in response to the protonation of a fully conserved histidine (His 193 of SoPIP2;1).

¹Department of Chemistry and Bioscience, Chalmers University of Technology, P O Box 462, SE-40530 Göteborg, Sweden. ²Theoretical and Computational Biophysics Group, Beckman Institute, University of Illinois at Urbana-Champaign, 405 N. Mathews Avenue, Urbana, Illinois 61801, USA. ³Department of Plant Biochemistry, Lund University, P O Box 124, SE-22100 Lund, Sweden.

be identified. We also determined the X-ray structure of SoPIP2;1 in its open conformation at 3.9 Å resolution, and performed molecular dynamics simulations to examine the initial events leading to the opening of the channel.

Crystal structure of SoPIP2;1

As with all other aquaporin crystals reported so far^{12–17}, SoPIP2;1 crystallized as a tetramer displaying extended hydrophobic interactions between monomers (Supplementary Fig. S3). In Fig. 2a the open (blue) and closed (green) X-ray structures of SoPIP2;1 are overlaid in stereo on that of bovine AQP0 (ref. 16) (light grey) and AQP1 (ref. 13) (grey). Although animal and plant evolutionary lines separated about 1.6 billion years ago²⁴, these highly conserved proteins have an identical structural core consistent with the 'hour-glass model'²⁵, differing only by 0.8 Å root-mean-square deviation on C α atoms within the transmembrane regions. Similarly, the half-helices formed by loops B (cytosolic) and E (extracellular), with the Asn-Pro-Ala (NPA) aquaporin signature motif at the amino-terminal ends, fold into the channel from opposite sides of the membrane and together create a seventh transmembrane region that is perfectly preserved structurally. Seven water molecules are observed within the SoPIP2;1 channel (Supplementary Fig. S4), revealing an unbroken water network stretching almost fully through the pore with a maximum distance of only 3.1 Å between each water molecule. During molecular dynamics simulations these

water molecules maintained a stable, continuous, single-file configuration and exhibited a highly correlated motion along the pore axis. As with all other simulated aquaporins^{18,19,26}, a specific re-orientation of water molecules across the NPA motif is observed, strongly indicating the presence of a positive electrostatic potential at the NPA region, which is suggested to prevent proton translocation in aquaporins^{27–31}.

Conformation of loop D

The most striking difference in the structure of SoPIP2;1 relative to the other aquaporins concerns the conformation of loop D (Fig. 2), which is typically four to seven residues longer for the PIP subfamily members than for other aquaporin homologues (Supplementary Fig. S2). In the closed conformation this extended loop folds underneath the aquaporin and occludes the water pore from accessing the cytosol. A key residue in this respect is the fully conserved Leu 197 of loop D, which inserts into a cavity near the entrance of the channel (Fig. 2a) and, in combination with His 99, Val 104 and Leu 108, creates a hydrophobic barrier blocking the pore. This structural feature is illustrated graphically in Fig. 3b because two water molecules separated by 6.4 Å are clearly visible on either side of the hydrophobic gate associated with Leu 197. Calculations of the pore diameter with the use of HOLE³² establish that, in the closed conformation, the pore narrows to a diameter of about 1.4 Å at Leu 197 and further narrows to 0.8 Å near Pro 195 and Val 194

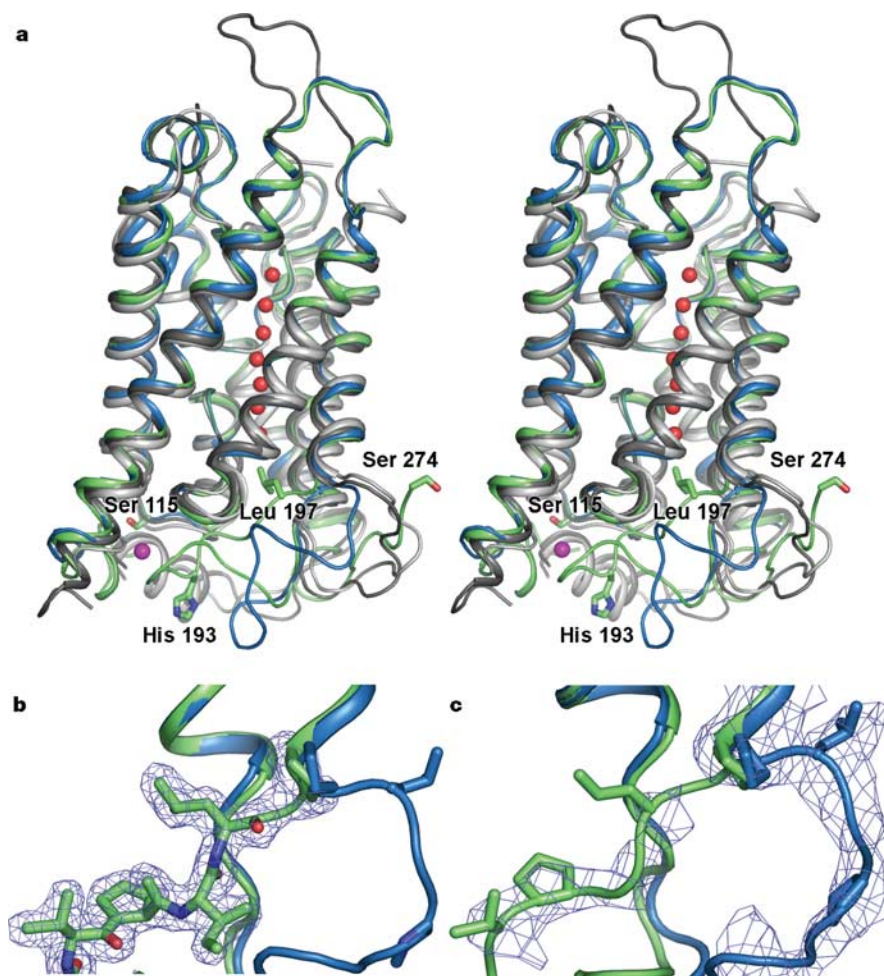


Figure 2 | Structures of the closed and open conformations of SoPIP2;1. **a**, Stereo models of SoPIP2;1 in its open (blue) and closed (green) conformations overlaid on that of AQP0 (light grey; Protein Data Bank (PDB) entry 1YMG) and AQP1 (grey; PDB entry 1J4N). **b**, **c**, Electron density for loop D in the closed (**b**) and open (**c**) conformations. The

structural model of the closed conformation is coloured green and the open conformation is coloured blue. Both $2F_o - F_c$ electron density maps are contoured at 1.0σ . Residual electron density in **c** indicates that the closed conformation is also present in partial occupancy.

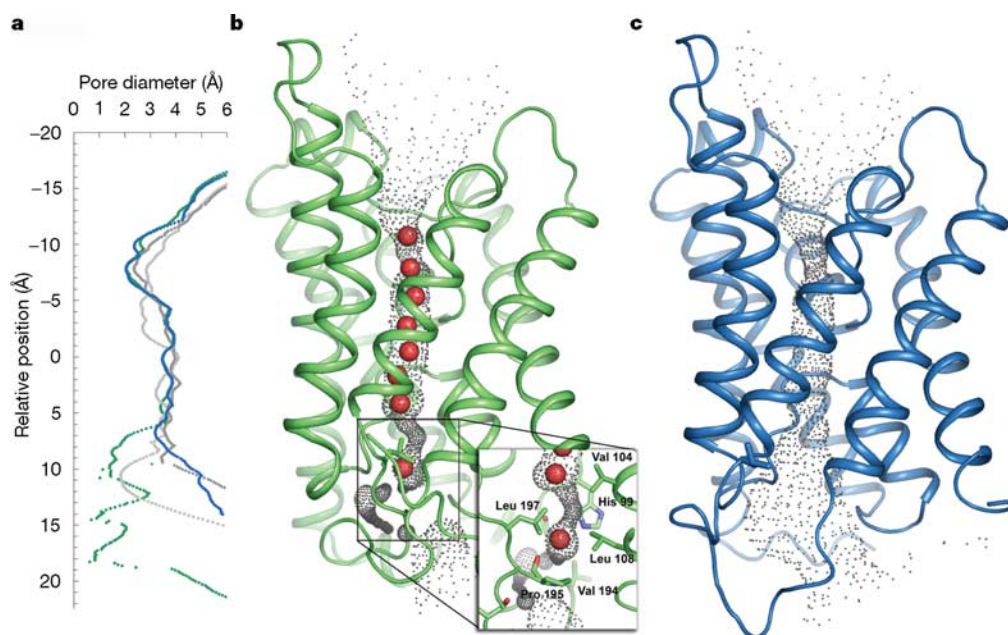


Figure 3 | Characterizing the SoPIP2;1 channel. **a**, The pore diameter of the closed conformation of SoPIP2;1 (green), and the open conformation of SoPIP2;1 (blue), AQP0 (light grey) and AQP1 (grey), represented as a function of the distance from the NPA signature sequence calculated with HOLE³². **b**, The same information for the closed conformation of SoPIP2;1

as in **a** but represented as a funnel illustrating the pore boundaries. The inset shows the pore near the gating region of loop D characterized by Leu 197, Pro 195 and Val 194. **c**, The same representation as in **b** but corresponding to the open conformation of SoPIP2;1.

(Fig. 3a, b). This compares with the minimum pore diameter of 2.1 Å within the SoPIP2;1 constriction region and severely restricts the passage of water. Molecular dynamics simulations of the non-phosphorylated protein also record a narrow pore (Fig. 4a) and a very low probability of finding a continuous water file in this region (Fig. 4b).

In the open structure of SoPIP2;1, by contrast, loop D is displaced up to 16 Å and the N terminus of helix 5 extends a further half-turn into the cytoplasm relative to that of its closed-conformation (Fig. 2, and Supplementary Fig. S5), although no hydrogen bond can form

between Leu 197 and Pro 201 in the usual $n + 4$ α -helical pattern. The open structure of SoPIP2;1 therefore becomes closer to that of AQP1 (ref. 13) (Figs 2a and 3a). Because of this extension of helix 5, the C α atoms of Leu 197, Pro 195 and Val 194 are displaced 8.4, 13.9 and 15.0 Å, respectively, away from the cytoplasmic pore entrance (Fig. 2, and Supplementary Fig. S5). Thus, by coupling the movement of loop D to the principal hydrophobic gate that blocks the channel, the pore diameter is locally increased to more than 4 Å (Fig. 3a, c) and the channel is open. A strong conformational coupling between loop D and the hydrophobic gate is clearly

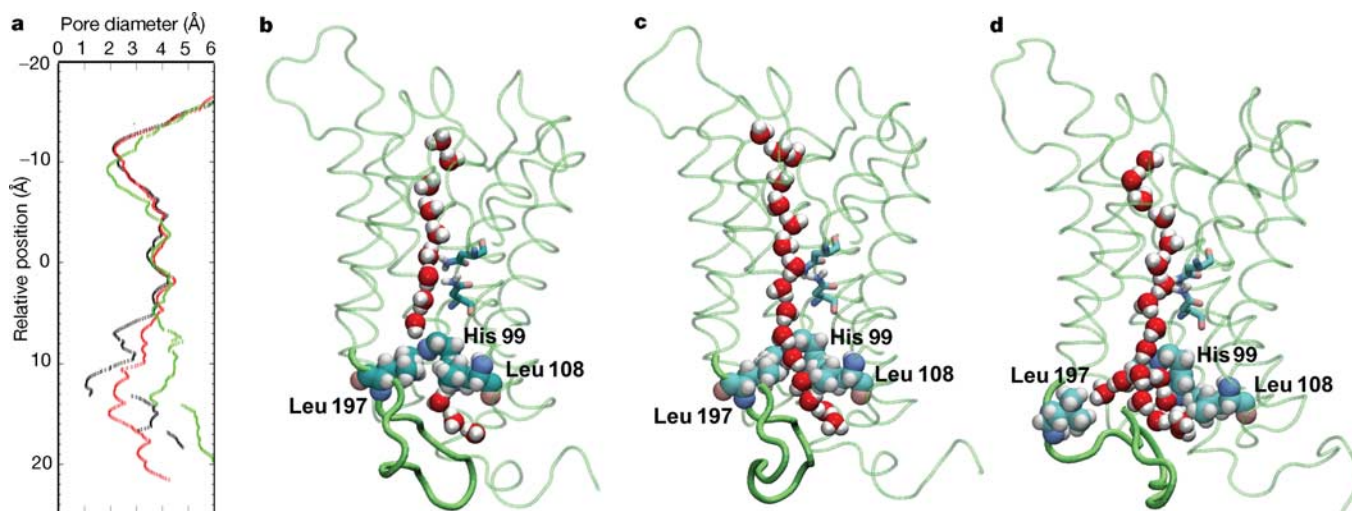


Figure 4 | Results of molecular dynamics simulations of the non-phosphorylated, phosphorylated, and induced open systems. **a**, The pore diameter of the non-phosphorylated (black), phosphorylated (red), and induced open (green) structures calculated with HOLE³², from structures averaged over the final 5 ns of each simulation. **b–d**, Snapshots of simulations of the non-phosphorylated (**b**), phosphorylated (**c**) and induced

open (**d**) structures. Loop D is drawn in a thicker tube representation to highlight its conformational coupling to the cytoplasmic gate. Residues forming the cytoplasmic gate are drawn in van der Waals representation. Two asparagine residues of the NPA motifs are shown in stick representation.

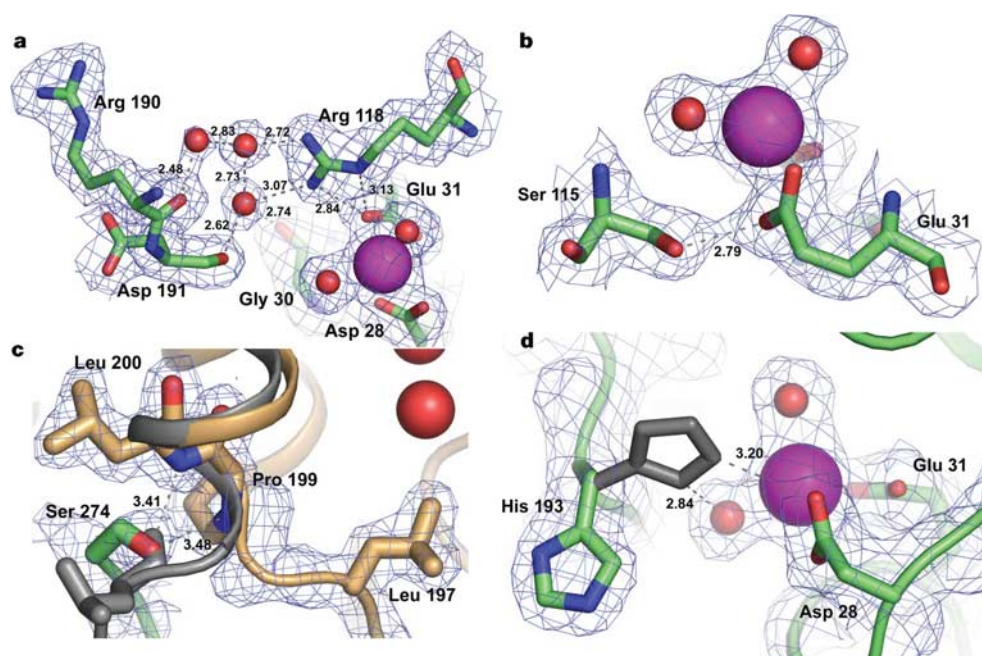


Figure 5 | Electron density at the sites of regulation by phosphorylation and pH for SoPIP2;1 in its closed conformation. **a**, Close-up view of the divalent-cation-binding site showing the location of the Cd^{2+} ion (purple) and the network of hydrogen bonds linking Gly 30 and Glu 31 through Arg 118 to Arg 190 and Asp 191 of loop D. **b**, Close-up view of the phosphorylation residue Ser 115, illustrating its hydrogen bond to Glu 31. **c**, Electron density for Ser 274, which contacts Pro 199 and Leu 200 of a

neighbouring monomer of the SoPIP2;1 tetramer. Overlaid in grey is the structure of the open conformation of SoPIP2;1, indicating that a steric clash with Leu 197 prevents helix 5 from adopting this conformation when Ser 274 is dephosphorylated. **d**, Close-up view of His 193. When protonated, an alternative conformation for His 193 (grey) may be adopted that forms a salt bridge to Asp 28. All $2F_o - F_c$ maps are contoured at 1.0σ . Numbers are distances in Å.

observed during molecular dynamics simulations (Fig. 4), in which a partial opening of the gate mediated through the displacement of loop D after phosphorylation of Ser 115 and Ser 274 was observed within 15 ns (Fig. 4a, c). It is also noteworthy that Tyr 149 of AQP0 (ref. 16), which has been suggested to gate AQP0 (refs 15, 33), overlays almost exactly with Leu 197 of SoPIP2;1 (Supplementary Fig. S6). For AQP0, however, only one amino-acid residue blocks the pore, whereas several residues of loop D cap the channel in SoPIP2;1.

Specific interactions governing gating

It has been postulated that divalent cations have a function in aquaporin regulation^{9,10} and inhibition^{33–35}. In Figs 2a and 5 a heavy metal (coloured purple; Supplementary Fig. S7 shows the anomalous difference density map) is observed near loop D and is assigned as Cd^{2+} because the addition of this ion improved the

crystal quality (see Supplementary Information). Cd^{2+} may be replaced by another divalent cation *in vivo*, and a search for similar structural motifs³⁶ revealed 13 PDB entries containing Ca^{2+} (Supplementary Information gives the accession codes). We therefore propose that this site binds Ca^{2+} *in vivo*. This divalent cation is implicated in regulation because it serves to anchor loop D, through a network involving ionic interactions and hydrogen bonds (Fig. 5a), onto a short α -helix of the N terminus and is critical for defining the closed conformation of SoPIP2;1. Specifically, Arg 190 and Asp 191 of loop D are connected to the side chain of Arg 118 (strictly conserved in PIPs) and Gly 30 through a hydrogen-bond network containing three water molecules. Arg 118 in turn forms hydrogen bonds to Glu 31 (strictly conserved in PIPs), which ligates the Cd^{2+} ion (Fig. 5a, b). Because loop D undergoes a large-scale rearrangement during the transition from the closed to the open conformation

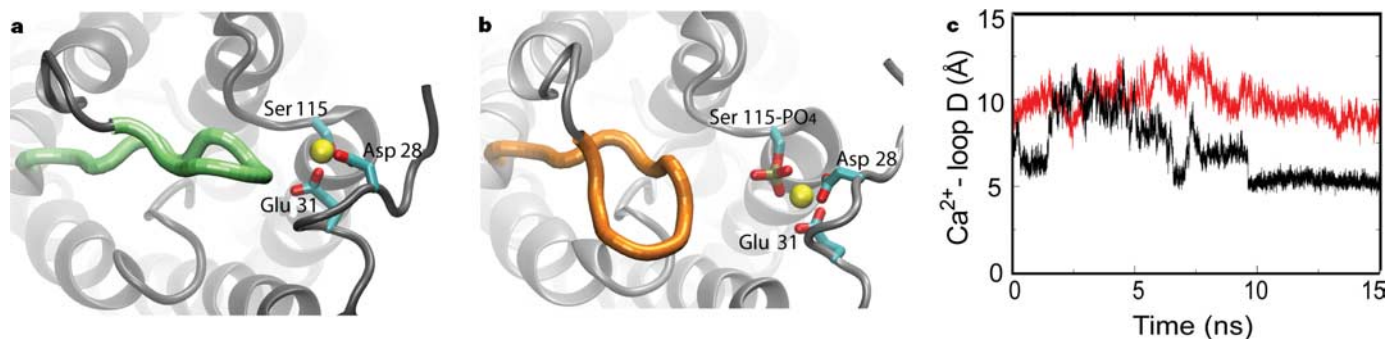


Figure 6 | Regulation of the cytoplasmic entrance into the SoPIP2;1 channel. **a**, **b**, Snapshots from non-phosphorylated (**a**) and phosphorylated (**b**) simulations. The Ca^{2+} ion is shown in yellow van der Waals representation and its coordinating residues (Ser 115, Asp 28 and Glu 31) are shown in stick representation. A mobile fraction of loop D (Ser 188 to

Ala 198) is shown in green (**a**) and orange (**b**) tube representation. **c**, The separation of the N terminus and loop D, defined as the distance between the Ca^{2+} ion and the closest C_α atom of loop D. Red, phosphorylated; black, non-phosphorylated.

of SoPIP2;1 (Fig. 2) it is apparent that the phosphorylation of Ser 115 disrupts this network of hydrogen bonds anchoring loop D to the N terminus. The issue of how this disruption evolves is analysed below with molecular dynamics simulations.

The regulatory mechanism of Ser 274 (conserved in PIP2 homologues) can also be understood from our structural results. In this case Ser 274 interacts with the backbone nitrogens of Pro 199 and Leu 200 of an adjacent monomer (Fig. 5c) because the C terminus of SoPIP2;1 extends towards the four-fold axis of the tetramer (Supplementary Fig. S3). From the X-ray structure of the open conformation we observe that, when these interactions are lost, the C terminus is disordered, and the position previously occupied by the side chain of Ser 274 becomes taken by the backbone carbonyl oxygen of Leu 197 (Fig. 5c, grey). Thus a steric clash is removed and the N terminus of helix 5 is able to extend an additional half-turn into the cytoplasm (Fig. 2), thereby displacing Leu 197 and opening the water channel (Fig. 3).

When the aquaporin is open, the protonation of His 193 (strictly conserved in PIPs) closes the channel¹¹. A mechanism for pH-regulated PIP gating is also apparent from the closed structure of SoPIP2;1. In Fig. 5d the conformation of His 193 is shown. At low pH, when His 193 is protonated, a simple rotation of the histidine side chain (Fig. 5d, grey) would enable it to form a salt bridge to Asp 28 (conserved in PIPs as Asp or Glu; Supplementary Fig. S2). In this manner the hydrogen-bond-mediated anchor for loop D onto the N terminus (Fig. 5a), which is lost when Ser 115 is phosphorylated, could be recovered. The structure of Ser 115-phosphorylated SoPIP2;1 at low pH can therefore be expected to be rather similar to that of the closed conformation reported here, with the cytoplasmic side of the aquaporin being capped by loop D, and Leu 197, Pro 195 and Val 194 effectively blocking the water channel.

Initial response to serine phosphorylation

To investigate the primary events that trigger the opening of the phosphorylated channel, molecular dynamics simulations of 15 ns duration were performed on the non-phosphorylated protein, and on the protein with both Ser 115 and Ser 274 phosphorylated, both starting from the closed crystal structure. We also conducted a 10-ns simulation starting from the phosphorylated closed structure during which loop D was forced (through constraints applied to residues 182 and 195–198 at the base of the loop; see Supplementary Information) to move towards its conformation in the open crystal structure. For all simulations Cd^{2+} was replaced with Ca^{2+} .

After 10 ns equilibration, the non-phosphorylated protein yielded a stable connection between loop D and the N terminus, as indicated by the short distance (about 5 Å; Fig. 6c) between this loop and the Ca^{2+} ion coordinated by residues in the N terminus (Fig. 6a). Furthermore, the side chains of residues 192–194 on loop D were observed to reorientate towards the N terminus and establish

direct and/or water-mediated hydrogen bond interactions with it, primarily through Glu 31. As is typically observed for loops interconnecting transmembrane helices, loop D is a flexible part of the protein and showed considerable fluctuations during the simulations, especially with regard to the position of its amino-acid side chains. Thus, all interactions between loop D and the N terminus observed during these simulations help to anchor this loop onto the cytoplasmic entrance into the pore (Supplementary Fig. S8, green conformation) and thereby decrease the probability of recovering an open channel.

During the 15-ns simulation of the phosphorylated enzyme, a significant rearrangement in the conformation of amino acids coordinating the Ca^{2+} ion occurred. This ion, which was (mainly) coordinated by Asp 28 and one side-chain oxygen of Glu 31 in the non-phosphorylated protein (Fig. 6a), became coordinated by both side-chain oxygens of Asp 28 and Glu 31 as well as the phosphate group of phosphorylated Ser 115 (Fig. 6b). This rearrangement resulted in an average displacement of the ion by 4.7 Å towards the periphery of the aquaporin tetramer in all monomers (Fig. 6b, c) and drew Glu 31 away from its original location, significantly disrupting the hydrogen-bond network connecting loop D to the N terminus. Thus released from its hydrogen-bond-mediated anchor, loop D underwent a significant conformational change away from the cytoplasmic entrance of the pore (Supplementary Fig. S8, orange conformation). Although the amplitude of the movement of loop D observed during this 15-ns simulation was considerably smaller than that observed for the X-ray structure of the open conformation, the nature of this movement (Supplementary Fig. S8, orange conformation) provides a plausible initial trajectory towards the open crystal conformation (Supplementary Fig. S8, blue conformation). We therefore suggest that the conformational change from the closed to the open structure proceeds first through the release of loop D from the N terminus, followed by the extension of the cytoplasmic end of helix 5 on a slower timescale (Fig. 2, and Supplementary Fig. S8).

A specific coupling between the movement of loop D and the opening of the hydrophobic gate near Leu 197 is also evident from simulations (Fig. 4). Before phosphorylation, this gate ($10 \text{ Å} < z < 15 \text{ Å}$; Fig. 4) completely blocks the access of the pore water to the cytoplasm (Fig. 4b) with the minimum diameter of the pore (averaged over the last 5 ns) being 1.04 Å as determined with HOLE³² (Fig. 4a, black line). After phosphorylation the pore diameter (averaged over the last 5 ns) was significantly increased in this region to 1.94 Å (Fig. 4a, red line) and a stable chain of water molecules connecting the inner channel to the cytoplasm was formed (Fig. 4c). When the loop of the phosphorylated protein was driven towards the loop conformation in the X-ray open structure, the cytoplasmic hydrophobic gate became completely open (average diameter of 3.54 Å over the last 5 ns; Fig. 4a, green line) and the cytoplasmic entrance to the channel became flooded with water (Fig. 4d).

Generality

A deeper understanding of the structural mechanism of water flow regulation provides insight into the unique physiology of plant cells. Structural results and simulations presented here reveal how the extended loop D of SoPIP2;1 containing Leu 197 provides a flexible lid, capping the channel from the cytoplasm and opening a hydrophobic gate after the phosphorylation of Ser 115 and Ser 274. This gating mechanism successfully unifies a significant body of biochemical^{9–11} and genetic evidence that has identified specific amino-acid residues governing plant aquaporin gating, and immediately suggests how the closed structure might be stabilized or destabilized. From the high degree of amino-acid sequence conservation this regulatory mechanism seems to be conserved in all plasma membrane aquaporins irrespective of plant species. Thus, the structural mechanism of plant aquaporin gating is applicable to

Table 1 | Crystallographic data and refinement statistics

	Closed conformation	Open conformation
Refinement		
Resolution (Å)	40.0–2.1	50.0–3.9
$R_{\text{work}}/R_{\text{free}}$	0.181/0.208	0.291/0.332
No. of atoms		
Protein	3756	6874
Ligand/ion	2	
Water	200	
B-factors		
Protein	29.7	69.0
Ligand/ion	30.6	
Water	39.6	
Root-mean-square deviations		
Bond lengths (Å)	0.02	0.024
Bond angles (°)	1.540	2.00

phenomena as everyday as a neglected pot-plant wilting and the flood irrigation of the rice paddies of Asia.

METHODS

Heterologous membrane protein overexpression. Functional SoPIP2;1, both as His-tagged and non-tagged protein, was overproduced in the methylotrophic yeast *Pichia pastoris*³⁷. Purification and concentration protocols are detailed in Supplementary Information.

Crystallization. Crystals of the closed conformation were obtained by hanging-drop vapour diffusion from the non-tagged protein. Sample (1 μ l) was mixed 1:1 with the reservoir solution containing 0.1 M Tris-HCl pH 8.0, 30% polyethylene glycol (PEG) 400 and 0.1 M NaCl. 0.1 M CdCl₂ was added to the drop in a 1:10 ratio. Crystallization setups were left to equilibrate at 4 °C and crystals appeared within a few days. Crystals of the open conformation were obtained by hanging-drop vapour diffusion from the His-tagged protein. Activity assays on the overexpressed His-tagged protein show significant water transport activity³⁷. Sample (1 μ l) was mixed 1:1 with the reservoir solution containing 0.1 M sodium citrate pH 5.6, 30% PEG 400 and 0.1 M NaCl. Crystallization setups were left to equilibrate at 4 °C and crystals appeared after six months.

Structure determination of the closed conformation. X-ray diffraction data (Table 1) from the closed conformation to 2.1 Å resolution were recorded at the Swiss Light Source (SLS) beamline X06SA, Villigen, Switzerland. Crystals belong to the space group I4, with two molecules in the asymmetric unit. Data were processed using MOSFLM and scaled using SCALA of the CCP4 suite³⁸. Molecular replacement was performed with the program MOLREP from the CCP4 program suite³⁸ with the coordinates of bovine AQP1 (PDB entry 1J4N; ref. 13) as the model. A clear solution with a correlation coefficient and an R-factor of 34.3% and 53.1%, respectively, was found. Automated model building was conducted by using ARP/WARP³⁹ and the resulting model was docked to the correct sequence using GUIDANCE³⁸. The model was subjected to multiple rounds of refinement in REFMAC5 (ref. 38), NCSREF³⁸ and CNS⁴⁰. The final R-factor and R_{free} are 18.1% and 20.8%, respectively. A single metal-binding site was unambiguously observed as a dominant peak in the anomalous difference density map (Supplementary Fig. S7) and was assigned as Cd²⁺ (see Supplementary Information). The quality of the structure was checked in PROCHECK⁴¹.

Structure determination of the open conformation. X-ray diffraction data (Table 1) from the open conformation to 3.9 Å resolution were collected at the European Synchrotron Radiation Facility (ESRF) beamline ID14 EH2, Grenoble, France. Crystals belong to the space group P2₁2₁2, with four molecules in the asymmetric unit. Molecular replacement was performed as above with the closed conformation as the starting model. A clear solution with a correlation coefficient and R-factor of 63.8% and 41.9%, respectively, was found. After the initial rigid-body refinement, only those residues that clearly showed distinct conformations from the closed structure were varied during structural refinement. The C terminus was disordered and could not be assigned. A new conformation for loop D was built into the composite omit 2F_o - F_c electron-density map in iterative steps. For each iteration the model was subjected to multiple rounds of refinement in CNS⁴⁰ with manual rebuilding in O⁴² between each round, and the composite omit map was recalculated. The R-factor and R_{free} are 29.1% and 33.2%, respectively. The quality of the structure was checked in PROCHECK⁴¹.

Molecular dynamics simulations. Molecular dynamics simulations were performed with tetrameric models of SoPIP2;1 embedded in a fully hydrated, 100 × 100-Å² POPE bilayer. Crystallographic water molecules were retained but all Cd²⁺ ions were replaced with Ca²⁺. The C termini (Ser 274) were capped with amide groups (CONH₂) to avoid artificially introducing a negative charge at this critical region. For all simulations the program NAMD2 (ref. 43) and the CHARMM27 parameter set^{44,45} were used. The Particle Mesh Ewald method⁴⁶ was employed for the computation of electrostatic forces. Langevin dynamics and a Langevin piston were used to maintain the temperature (300 K) and pressure (1 atm), respectively. Model building and analysis were done with VMD⁴⁷. Three systems were simulated: first, the unphosphorylated system, representing a 'closed' channel; second, the 'phosphorylated' channel, in which both Ser 115 and Ser 274 were phosphorylated to trigger the opening; and third, the 'induced open' channel in which C α coordinates of residues 182 and 195–198 in the open crystal structure were used to induce a fully open structure. After energy minimization and a 500-ps protein-fixed simulation, systems 1 and 2 each were simulated for 15 ns. In system 3, the system was subjected to an additional constrained simulation phase, in which position constraints were applied to C α atoms of residues 182 and 195–198 in seven 100-ps steps with varying force constants to induce an open structure. After the initial 700 ps of constrained simulation, all constraints were removed and the system was simulated for 10 ns. More details are provided in Supplementary Information.

Received 16 February; accepted 14 October 2005.

Published online 7 December 2005.

- Preston, G. M., Carroll, T. P., Guggino, W. B. & Agre, P. Appearance of water channels in *Xenopus* oocytes expressing red cell CHIP28 protein. *Science* **256**, 385–387 (1992).
- Johansson, I., Karlsson, M., Johanson, U., Larsson, C. & Kjellbom, P. The role of aquaporins in cellular and whole plant water balance. *Biochim. Biophys. Acta* **1465**, 324–342 (2000).
- Agre, P. & Kozono, D. Aquaporin water channels: molecular mechanisms for human diseases. *FEBS Lett.* **555**, 72–78 (2003).
- King, L. S., Kozono, D. & Agre, P. From structure to disease: the evolving tale of aquaporin biology. *Nature Rev. Mol. Cell Biol.* **5**, 687–698 (2004).
- Tamas, M. J. et al. A short regulatory domain restricts glycerol transport through yeast Fps1p. *J. Biol. Chem.* **278**, 6337–6345 (2003).
- Johanson, U. et al. The complete set of genes encoding major intrinsic proteins in *Arabidopsis* provides a framework for a new nomenclature for major intrinsic proteins in plants. *Plant Physiol.* **126**, 1358–1369 (2001).
- Morishita, Y., Sakube, Y., Sasaki, S. & Ishibashi, K. Molecular mechanisms and drug development in aquaporin water channel diseases: aquaporin superfamily (superaquaporins): expansion of aquaporins restricted to multicellular organisms. *J. Pharmacol. Sci.* **96**, 276–279 (2004).
- Borstlap, A. C. Early diversification of plant aquaporins. *Trends Plant Sci.* **7**, 529–530 (2002).
- Johansson, I. et al. Water transport activity of the plasma membrane aquaporin PM28A is regulated by phosphorylation. *Plant Cell* **10**, 451–459 (1998).
- Johansson, I., Larsson, C., Ek, B. & Kjellbom, P. The major integral proteins of spinach leaf plasma membranes are putative aquaporins and are phosphorylated in response to Ca²⁺ and apoplastic water potential. *Plant Cell* **8**, 1181–1191 (1996).
- Tournaire-Roux, C. et al. Cytosolic pH regulates root water transport during anoxic stress through gating of aquaporins. *Nature* **425**, 393–397 (2003).
- Murata, K. et al. Structural determinants of water permeation through aquaporin-1. *Nature* **407**, 599–605 (2000).
- Sui, H., Han, B. G., Lee, J. K., Walian, P. & Jap, B. K. Structural basis of water-specific transport through the AQP1 water channel. *Nature* **414**, 872–878 (2001).
- Savage, D. F., Egea, P. F., Robles-Colmenares, Y., O'Connell, J. D. & Stroud, R. M. Architecture and selectivity in aquaporins: 2.5 Å X-ray structure of aquaporin Z. *PLoS Biol.* **1**, E72 (2003).
- Gonen, T., Sliz, P., Kistler, J., Cheng, Y. & Walz, T. Aquaporin-0 membrane junctions reveal the structure of a closed water pore. *Nature* **429**, 193–197 (2004).
- Harries, W. E., Akhavan, D., Miercke, L. J., Khademi, S. & Stroud, R. M. The channel architecture of aquaporin 0 at a 2.2 Å resolution. *Proc. Natl Acad. Sci. USA* **101**, 14045–14050 (2004).
- Fu, D. et al. Structure of a glycerol-conducting channel and the basis for its selectivity. *Science* **290**, 481–486 (2000).
- Tajkhorshid, E. et al. Control of the selectivity of the aquaporin water channel family by global orientational tuning. *Science* **296**, 525–530 (2002).
- de Groot, B. L. & Grubmüller, H. Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and GlpF. *Science* **294**, 2353–2357 (2001).
- Jensen, M. O., Tajkhorshid, E. & Schulten, K. The mechanism of glycerol conduction in aquaglyceroporins. *Structure* **9**, 1083–1093 (2001).
- Jensen, M. O., Park, S., Tajkhorshid, E. & Schulten, K. Energetics of glycerol conduction through aquaglyceroporin GlpF. *Proc. Natl Acad. Sci. USA* **99**, 6731–6736 (2002).
- Daniels, M. J., Chrispeels, M. J. & Yeager, M. Projection structure of a plant vacuole membrane aquaporin by electron cryo-crystallography. *J. Mol. Biol.* **294**, 1337–1349 (1999).
- Kukulski, W. et al. The 5 Å structure of heterologously expressed plant aquaporin SoPIP2;1. *J. Mol. Biol.* **350**, 611–616 (2005).
- Hedges, S. B., Blair, J. E., Venturi, M. L. & Shoe, J. L. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol.* **4**, 2 (2004).
- Jung, J. S., Preston, G. M., Smith, B. L., Guggino, W. B. & Agre, P. Molecular structure of the water channel through aquaporin CHIP. The hourglass model. *J. Biol. Chem.* **269**, 14648–14654 (1994).
- Wang, Y., Schulten, K. & Tajkhorshid, E. What makes an aquaporin a glycerol channel: A comparative study of AqpZ and GlpF. *Structure* **13**, 1107–1118 (2005).
- de Groot, B. L., Frigato, T., Helms, V. & Grubmüller, H. The mechanism of proton exclusion in the aquaporin-1 water channel. *J. Mol. Biol.* **333**, 279–293 (2003).
- Jensen, M. O., Tajkhorshid, E. & Schulten, K. Electrostatic tuning of permeation and selectivity in aquaporin water channels. *Biophys. J.* **85**, 2884–2899 (2003).
- Chakrabarti, N., Tajkhorshid, E., Roux, B. & Pomes, R. Molecular basis of proton blockage in aquaporins. *Structure* **12**, 65–74 (2004).
- Ilán, B., Tajkhorshid, E., Schulten, K. & Voth, G. A. The mechanism of proton exclusion in aquaporin channels. *Proteins* **55**, 223–228 (2004).

31. de Groot, B. L. & Grubmüller, H. The dynamics and energetics of water permeation and proton exclusion in aquaporins. *Curr. Opin. Struct. Biol.* **15**, 176–183 (2005).
32. Smart, O. S., Goodfellow, J. M. & Wallace, B. A. The pore dimensions of gramicidin A. *Biophys. J.* **65**, 2455–2460 (1993).
33. Nemeth-Cahalan, K. L. & Hall, J. E. pH and calcium regulate the water permeability of aquaporin O. *J. Biol. Chem.* **275**, 6777–6782 (2000).
34. Zelenina, M., Bondar, A. A., Zelenin, S. & Aperia, A. Nickel and extracellular acidification inhibit the water permeability of human aquaporin-3 in lung epithelial cells. *J. Biol. Chem.* **278**, 30037–30043 (2003).
35. Zelenina, M., Tritto, S., Bondar, A. A., Zelenin, S. & Aperia, A. Copper inhibits the water and glycerol permeability of aquaporin-3. *J. Biol. Chem.* **279**, 51939–51943 (2004).
36. Madsen, D. & Kleywegt, G. J. Interactive motif and fold recognition in protein structures. *J. Appl. Crystallogr.* **35**, 137–139 (2001).
37. Karlsson, M. *et al.* Reconstitution of water channel function of an aquaporin overexpressed and purified from *Pichia pastoris*. *FEBS Lett.* **537**, 68–72 (2003).
38. Bailey, S. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50**, 760–763 (1994).
39. Morris, R. J., Perrakis, A. & Lamzin, V. S. ARP/wARP and automatic interpretation of protein electron density maps. *Methods Enzymol.* **374**, 229–244 (2003).
40. Brunger, A. T. *et al.* Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
41. Laskowski, R. A., Rullmann, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* **8**, 477–486 (1996).
42. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
43. Kalé, L. *et al.* NAMD2: Greater scalability for parallel molecular dynamics. *J. Comp. Phys.* **151**, 283–312 (1999).
44. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
45. Schlenkerich, M., Brickmann, J., MacKerell, A. D. & Karplus, M. in *Biological Membranes: A Molecular Perspective from Computation and Experiment* (eds Merz, K. M. & Roux, B.) 31–81 (Birkhauser, Boston, Massachusetts, 1996).
46. Darden, T., York, D. & Pedersen, L. Particle Mesh Ewald—an $N \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
47. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Larsson for useful discussions; the Pittsburgh Supercomputer Center and the National Center for Supercomputing Applications for providing computer time; and the European Synchrotron Radiation Facility and the Swiss Light Source for access to synchrotron radiation. Financial support was provided by Formas, the Research School of Pharmaceutical Sciences (FLÄK), Swegene, the Swedish Research Council (VR), the Swedish Strategic Research Foundation (SSF), the European Commission Integrated Projects EMEP and SPINE, the Chalmers Bioscience Programme and the NIH.

Author Information The coordinates and structure factor amplitudes for the closed and open structures have been deposited in the Protein Data Bank under the accession codes 1Z98 and 2B5F, respectively. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.N. (richard.neutze@chembio.chalmers.se), P.K. (per.kjellbom@plantbio.lu.se) or E.T. (emad@ks.uiuc.edu).

Discovery of very-high-energy γ -rays from the Galactic Centre ridge

F. Aharonian¹, A. G. Akhperjanian², A. R. Bazer-Bachi³, M. Beilicke⁴, W. Benbow¹, D. Berge¹, K. Bernlöhr^{1,5}, C. Boisson⁶, O. Bolz¹, V. Borrel³, I. Braun¹, F. Breitling⁵, A. M. Brown⁷, P. M. Chadwick⁷, L.-M. Chounet⁸, R. Cornils⁴, L. Costamante^{1,20}, B. Degrange⁸, H. J. Dickinson⁷, A. Djannati-Atai⁹, L. O'C. Drury¹⁰, G. Dubus⁸, D. Emmanoulopoulos¹¹, P. Espigat⁹, F. Feinstein¹², G. Fontaine⁸, Y. Fuchs¹³, S. Funk¹, Y. A. Gallant¹², B. Giebels⁸, S. Gillessen¹, J. F. Glicenstein¹⁴, P. Goret¹⁴, C. Hadjichristidis⁷, D. Hauser¹, M. Hauser¹¹, G. Heinzelmann⁴, G. Henri¹³, G. Hermann¹, J. A. Hinton¹, W. Hofmann¹, M. Holleran¹⁵, D. Horns¹, A. Jacholkowska¹², O. C. de Jager¹⁵, B. Khélifi¹, S. Klages¹, Nu. Komin⁵, A. Konopelko⁵, I. J. Latham⁷, R. Le Gallou⁷, A. Lemièr⁹, M. Lemoine-Goumard⁸, N. Leroy⁸, T. Lohse⁵, A. Marcowith³, J. M. Martin⁶, O. Martineau-Huynh¹⁶, C. Masterson^{1,20}, T. J. L. McComb⁷, M. de Naurois¹⁶, S. J. Nolan⁷, A. Noutsos⁷, K. J. Orford⁷, J. L. Osborne⁷, M. Ouchrif^{16,20}, M. Panter¹, G. Pelletier¹³, S. Pita⁹, G. Pühlhofer¹¹, M. Punch⁹, B. C. Raubenheimer¹⁵, M. Raue⁴, J. Raux¹⁶, S. M. Rayner⁷, A. Reimer¹⁷, O. Reimer¹⁷, J. Ripken⁴, L. Rob¹⁸, L. Rolland¹⁶, G. Rowell¹, V. Sahakian², L. Saugé¹³, S. Schlenker⁵, R. Schlickeiser¹⁷, C. Schuster¹⁷, U. Schwanke⁵, M. Siewert¹⁷, H. Sol⁶, D. Spangler⁷, R. Steenkamp¹⁹, C. Stegmann⁵, J.-P. Tavernet¹⁶, R. Terrier⁹, C. G. Théoret⁹, M. Tluczykont^{8,20}, C. van Eldik¹, G. Vasileiadis¹², C. Venter¹⁵, P. Vincent¹⁶, H. J. Völk¹ & S. J. Wagner¹¹

The source of Galactic cosmic rays (with energies up to 10^{15} eV) remains unclear, although it is widely believed that they originate in the shock waves of expanding supernova remnants^{1,2}. At present the best way to investigate their acceleration and propagation is by observing the γ -rays produced when cosmic rays interact with interstellar gas³. Here we report observations of an extended region of very-high-energy ($>10^{11}$ eV) γ -ray emission correlated spatially with a complex of giant molecular clouds in the central 200 parsecs of the Milky Way. The hardness of the γ -ray spectrum and the conditions in those molecular clouds indicate that the cosmic rays giving rise to the γ -rays are likely to be protons and nuclei rather than electrons. The energy associated with the cosmic rays could have come from a single supernova explosion around 10^4 years ago.

The observations described here were carried out with the High Energy Stereoscopic System (HESS), a system of four imaging atmospheric-Cherenkov telescopes⁴. The instrument operates in the teraelectronvolt energy range (TeV), beyond the regime accessible to satellite-based detectors (MeV up to ~ 10 GeV). At satellite energies, the technique of probing the distribution of cosmic rays using γ -ray emission has been demonstrated in the large-scale mapping of the Galactic plane by EGRET⁵. The γ -ray flux was found to approximately trace the density of interstellar gas, illustrating that the flux of cosmic rays is roughly constant throughout the Galaxy. However, given its modest angular resolution ($\sim 1^\circ$),

EGRET could only resolve the few nearest molecular clouds. The order-of-magnitude better angular resolution of HESS opens up this possibility of resolving individual clouds out to the distance of the Galactic Centre. Moreover, in the energy range accessible to EGRET, the picture is complicated owing to three comparable contributions to the diffuse γ -ray flux (from cosmic ray electrons¹ via inverse Compton scattering and Bremsstrahlung radiation and from cosmic ray protons via neutral pion decay). In the energy range of HESS the dominant component of the truly diffuse γ -ray emission is very probably the decay of neutral pions produced in the interactions of cosmic rays with ambient material. Taken together, the wide field of view ($\sim 5^\circ$) and the improved angular resolution (better than 0.1°) of HESS have made possible the mapping of extended γ -ray emission.

Early HESS observations of the Galactic Centre region led to the detection of a point-like source of very-high-energy (VHE) γ -rays at the gravitational centre of the Galaxy (HESS J1745–290) (ref. 6), compatible with the positions of the supermassive black hole Sagittarius A*, the supernova remnant Sagittarius A East, and a Galactic Centre source reported by other groups^{7,8}. A more sensitive exposure of the region in 2004 revealed a second source: the supernova remnant/pulsar wind nebula G0.9+0.1 (ref. 9). These two sources are clearly visible in Fig. 1a. For previous VHE instruments such sources were close to the limit of detectability. With the greater sensitivity of the HESS instrument it is possible to subtract these two sources and search for much fainter emission. Subtracting the best-fit

¹Max-Planck Institut für Kernphysik, PO Box 103980, D 69029 Heidelberg, Germany. ²Yerevan Physics Institute, 2 Alikhanian Brothers St., 375036 Yerevan, Armenia. ³Centre d'Etude Spatiale des Rayonnements, CNRS/UPS, 9 av. du Colonel Roche, BP 4346, F-31029 Toulouse Cedex 4, France. ⁴Universität Hamburg, Institut für Experimentalphysik, Luruper Chaussee 149, D 22761 Hamburg, Germany. ⁵Institut für Physik, Humboldt-Universität zu Berlin, Newtonstr. 15, D 12489 Berlin, Germany. ⁶LUTH, UMR 8102 du CNRS, Observatoire de Paris, Section de Meudon, F-92195 Meudon Cedex, France. ⁷University of Durham, Department of Physics, South Road, Durham DH1 3LE, UK. ⁸Laboratoire Leprince-Ringuet, IN2P3/CNRS, Ecole Polytechnique, F-91128 Palaiseau, France. ⁹APC, 11 Place Marcelin Berthelot, F-75231 Paris Cedex 05, France (UMR 7164-CNRS, Université Paris VII, CEA, Observatoire de Paris). ¹⁰Dublin Institute for Advanced Studies, 5 Merrion Square, Dublin 2, Ireland. ¹¹Landessternwarte, Königstuhl, D 69117 Heidelberg, Germany. ¹²Laboratoire de Physique Théorique et Astroparticules, IN2P3/CNRS, Université Montpellier II, CC 70, Place Eugène Bataillon, F-34095 Montpellier Cedex 5, France. ¹³Laboratoire d'Astrophysique de Grenoble, INSU/CNRS, Université Joseph Fourier, BP 53, F-38041 Grenoble Cedex 9, France. ¹⁴DAPNIA/DSM/CEA, CE Saclay, F-91191 Gif-sur-Yvette, Cedex, France. ¹⁵Unit for Space Physics, North-West University, Potchefstroom 2520, South Africa. ¹⁶Laboratoire de Physique Nucléaire et de Hautes Energies, IN2P3/CNRS, Universités Paris VI & VII, 4 Place Jussieu, F-75252 Paris Cedex 5, France. ¹⁷Institut für Theoretische Physik, Lehrstuhl IV: Weltraum und Astrophysik, Ruhr-Universität Bochum, D 44780 Bochum, Germany. ¹⁸Institute of Particle and Nuclear Physics, Charles University, V Holesovickach 2, 180 00 Prague 8, Czech Republic. ¹⁹University of Namibia, Private Bag 13301, Windhoek, Namibia. ²⁰European Associated Laboratory for Gamma-Ray Astronomy.

model for point-like emission at the position of these excesses yields the map shown in Fig. 1b. Two significant features are apparent after subtraction: extended emission spatially coincident with the unidentified EGRET source 3EG J1744–3011 (discussed in ref. 10) and emission extending along the Galactic plane for roughly 2° . The latter emission is not only very clearly extended in longitude l , but also significantly extended in latitude b (beyond the angular resolution of HESS) with a characteristic root-mean-square (r.m.s.) width of 0.2° , as can be seen in the Galactic latitude slices shown in Fig. 2. The reconstructed γ -ray spectrum for the region $|l| < 0.8^\circ$, $|b| < 0.3^\circ$ (with point-source emission subtracted) is well described by a power law with photon index $\Gamma = 2.29 \pm 0.07_{\text{stat}} \pm 0.20_{\text{sys}}$ (Fig. 3; see the Supplementary Information for a discussion of systematic errors).

Given the plausible assumption that the γ -ray emission takes place near the centre of the Galaxy, at a distance of about 8.5 kpc, the observed r.m.s. extension in latitude of 0.2° corresponds to a scale of ~ 30 pc. This value is similar to that of interstellar material in giant

molecular clouds in this region, as traced by their CO emission and in particular by their CS emission¹¹. CS line emission does not suffer from the problem of ‘standard’ CO lines¹²: that clouds are optically thick for these lines and hence the total mass of clouds may be underestimated. The CS data suggest that the central region of the Galaxy, $|l| < 1.5^\circ$ and $|b| < 0.25^\circ$, contains about $3\text{--}8 \times 10^7$ solar masses of interstellar gas, structured in a number of overlapping clouds, which provide an efficient target for the nucleonic cosmic rays permeating these clouds. The region over which the γ -ray spectrum is integrated contains 55% of the CS emission corresponding to a mass of $1.7\text{--}4.4 \times 10^7$ solar masses. At least for $|l| < 1^\circ$, we find a close match between the distribution of the VHE γ -ray emission and the density of dense interstellar gas as traced by CS emission (Fig. 1b and Fig. 2).

The close correlation between γ -ray emission and available target material in the central 200 pc of our galaxy is a strong indication for an origin of this emission in the interactions of cosmic rays. Following this interpretation, the similarity in the distributions of CS line and VHE γ -ray emission implies a rather uniform CR density in the region. In the case of a power-law energy distribution the spectral index of the γ -rays closely traces the spectral index of the cosmic rays themselves (corrections due to scaling violations in the cosmic-ray interactions are small, $\Delta\Gamma < 0.1$; see Supplementary Information), so the measured γ -ray spectrum implies a cosmic-ray spectrum near the Galactic Centre with a spectral index close to 2.3, significantly harder than in the solar neighbourhood (where an index of 2.75 is measured). Given the probable proximity of particle accelerators, propagation effects are likely to be less pronounced than in the Galaxy as a whole, providing a natural explanation for the harder spectrum which is closer to the intrinsic cosmic-ray-source spectra. The main uncertainty in estimating the flux of cosmic rays in the Galactic Centre is the uncertainty in the amount of target material. Following ref. 3 and using the mass estimate of ref. 11 we can estimate the expected γ -ray flux from the region, assuming for the moment that the Galactic Centre cosmic-ray flux and spectrum are identical to those measured in the solar neighbourhood. Figure 3 shows the expected γ -ray flux as a grey band, together with the observed spectrum. While below 500 GeV there is reasonable agreement with this simple prediction, there are clearly more high-energy γ -rays than expected. The γ -ray flux above 1 TeV is a factor of 3–9 higher than the expected flux. The implication is that the number density of cosmic rays with multi-TeV energies exceeds the local density by the same factor. The size of the enhancement increases rapidly at energies above 1 TeV.

The observation of correlation between target material and TeV γ -ray emission is unique and provides a compelling case for an origin of the emission in the interactions of cosmic-ray nuclei. In addition, the harder-than-expected spectrum and the higher-than-expected TeV flux imply that there is an additional component to the Galactic Centre cosmic-ray population above the cosmic-ray ‘sea’ that fills the Galaxy. This is the first time that such direct evidence for recently accelerated (hadronic) cosmic rays in any part of our Galaxy has been found. The energy required to accelerate this additional component is estimated to be 10^{49} erg in the energy range 4–40 TeV or $\sim 10^{50}$ erg in total if the measured spectrum extends from 10^9 – 10^{15} eV. Given a typical supernova explosion energy of 10^{51} erg, the observed cosmic ray excess could have been produced in a single supernova remnant, assuming a 10% efficiency for cosmic-ray acceleration. In such a scenario, any epoch of cosmic-ray production must have occurred in the recent enough past that the rays that were accelerated have not yet diffused out of the Galactic Centre region. Representing the diffusion of protons with energies of several TeV in the form $D = \eta \times 10^{30} \text{ cm}^2 \text{ s}^{-1}$ (where $10^{30} \text{ cm}^2 \text{ s}^{-1}$ is the approximate value of the diffusion coefficient in the Galactic disk at TeV energies), we estimate the diffusion timescale to be $t = R^2/2D \approx 3,000(\theta/1^\circ)^2/\eta$ years, where θ is the angular distance from the Galactic Centre. Owing to the larger magnetic field and higher turbulence in the

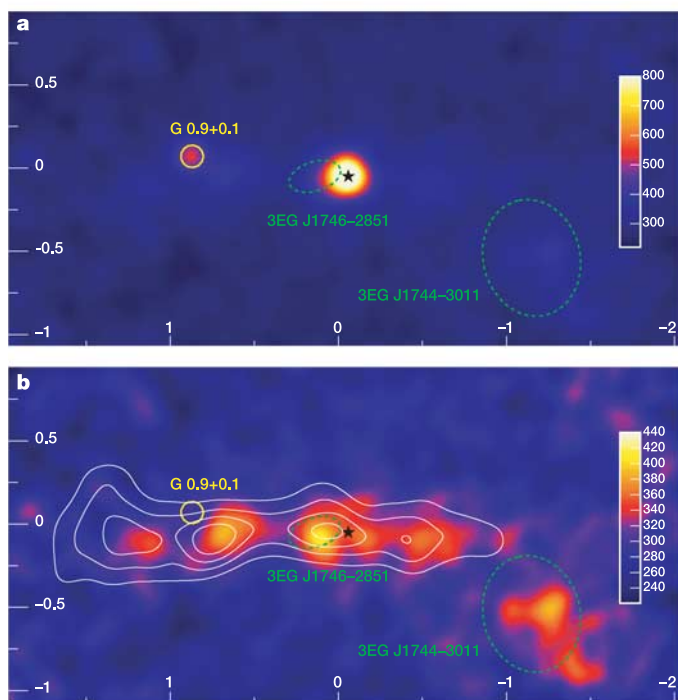


Figure 1 | VHE γ -ray images of the Galactic Centre region. **a**, γ -ray count map; **b**, the same map after subtraction of the two dominant point sources, showing an extended band of gamma-ray emission. Axes are Galactic latitude (x) and Galactic longitude (y), units are degrees. The colour scale is in ‘events’ and is dimensionless. White contour lines indicate the density of molecular gas, traced by its CS emission. The position and size of the composite supernova remnant G0.9+0.1 is shown with a yellow circle. The position of Sgr A* is marked with a black star. The 95% confidence region for the positions of the two unidentified EGRET sources in the region are shown as dashed green ellipses²⁰. These smoothed and acceptance-corrected images are derived from 55 hours of data consisting of dedicated observations of Sgr A*, G0.9+0.1 and a part of the data of the HESS Galactic plane survey²¹. The excess observed along the Galactic plane consists of $\sim 3,500$ γ -ray photons and has a statistical significance of 14.6 standard deviations. The absence of any residual emission at the position of the point-like γ -ray source G0.9+0.1 demonstrates the validity of the subtraction technique. The energy threshold of the maps is 380 GeV, owing to the tight γ -ray selection cuts applied here to improve signal/noise and angular resolution. We note that the ability of HESS to map extended γ -ray emission has been demonstrated for the shell-type supernova remnants RXJ1713.7–3946 (ref. 22) and RX J0852.0–4622 (ref. 23). The white contours are evenly spaced and show velocity integrated CS line emission from ref. 11, and have been smoothed to match the angular resolution of HESS.

central region compared to more conventional regions of the Galactic disk, the normalization parameter η is probably ≤ 1 and a source or sources of age ~ 10 kyr could fill the region $|l| < 1^\circ$ with cosmic rays. Indeed, the observation of a deficit in VHE emission at $l = 1.3^\circ$ relative to the available target material (see Fig. 2) suggests that cosmic rays, which were recently accelerated in a source or sources in the Galactic Centre region, have not yet diffused out beyond $|l| = 1^\circ$.

The observed morphology and spectrum of the γ -ray emission provide evidence that one or more cosmic-ray accelerators have been active in the Galactic Centre in the last 10,000 years. The fact that the diffuse emission exhibits a photon index Γ that is the same—within errors—as that of the central source HESS J1745–290 suggests that this object could be the source in question. Within the $1'$ error box of HESS J1745–290 are two compelling candidates for such a cosmic-ray accelerator. The first is the supernova remnant Sgr A East (ref. 13)

with its estimated age around 10 kyr (ref. 14); younger ages have been quoted for Sgr A East (ref. 15), reflecting the significant uncertainty in this estimate. The second is the supermassive black hole Sgr A* (refs 16, 17), which may have been more active in the past.

Another alternative possibility is that a population of electron accelerators produces the observed γ -ray emission via inverse Compton scattering. Extended objects with photon indices close to the value 2.3 observed in the Galactic Centre are observed elsewhere in the Galactic plane¹⁰. The parent population of objects such as pulsar wind nebulae (that is, massive stars) would probably approximately follow the distribution of molecular gas. However, in the intense photon fields and high magnetic fields within and close to the Galactic Centre molecular clouds^{18,19}, TeV electrons would lose their energy rapidly: $t_{\text{rad}} \approx 120(B/100 \mu\text{G})^{-2}(E_e/10 \text{ TeV})^{-1}$ years. We would therefore expect to see rather compact sources (point-like

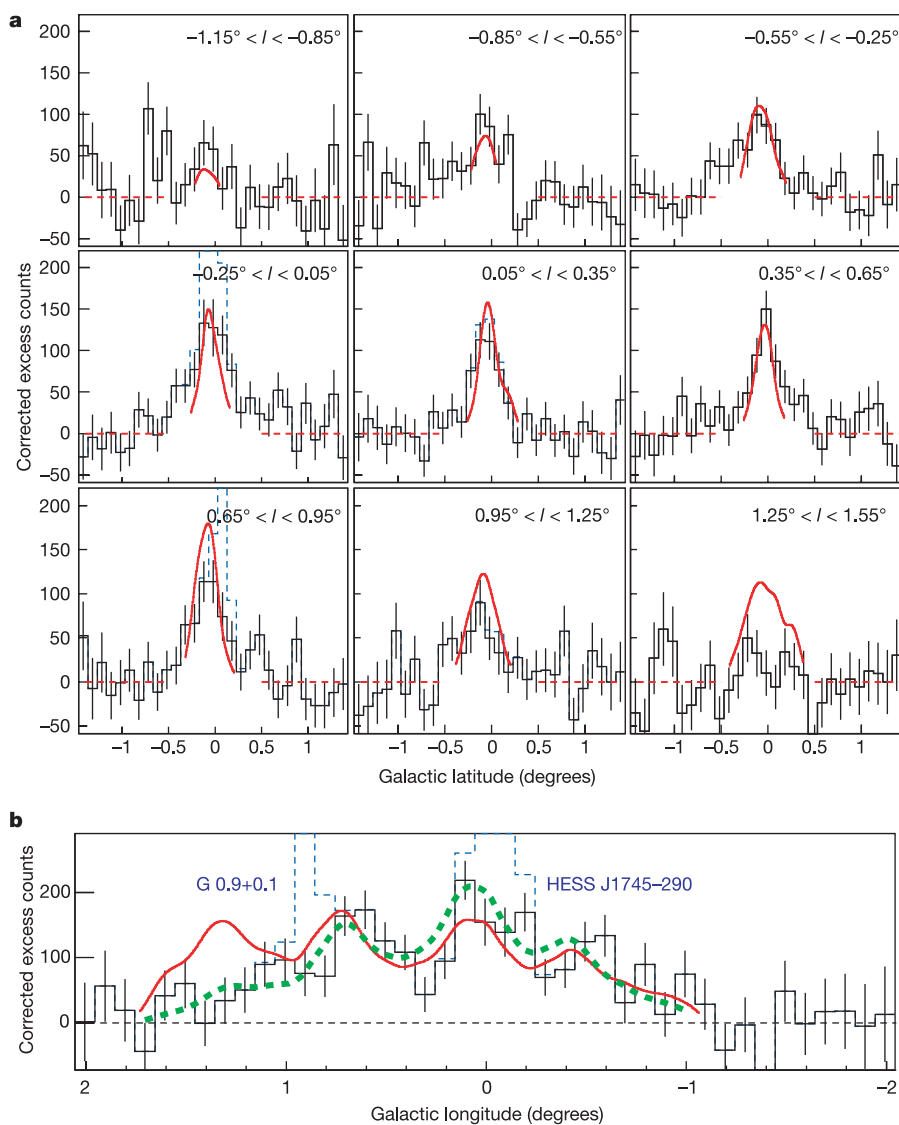


Figure 2 | γ -ray emission from the Galactic Centre region. Distribution of γ -ray emission in Galactic latitude (for individual slice in longitude, **a**) and in Galactic longitude (**b**). The red curves show the density of molecular gas, traced by CS emission. The upper panel shows acceptance-corrected (and cosmic-ray background-subtracted) γ -ray counts for 0.3° -wide bands in longitude. The point-source-subtracted counts are shown in black. The dashed blue histogram shows the unsmoothed values (the y scale is truncated). The red curves correspond to the smoothed CS map of Fig. 1 and are drawn only in the regions where CS measurements are available. The

dashed red lines show nominal zero CS density in regions away from the Galactic plane. Panel **b** shows γ -ray counts versus l for $-0.2^\circ < b < 0.2^\circ$. The CS line flux may be underestimated close to $l = -1^\circ$, owing to a narrower coverage in b at this longitude. The dashed line shows the γ -ray flux expected if the cosmic-ray density distribution can be described by a gaussian centred at $l = 0^\circ$ and with r.m.s. 0.8° , as expected in a simple model for diffusion away from a central source of age $\sim 10^4$ years. In all plots the background level is estimated using events from the regions $0.8^\circ < |b| < 1.5^\circ$. Error bars show ± 1 standard deviation.

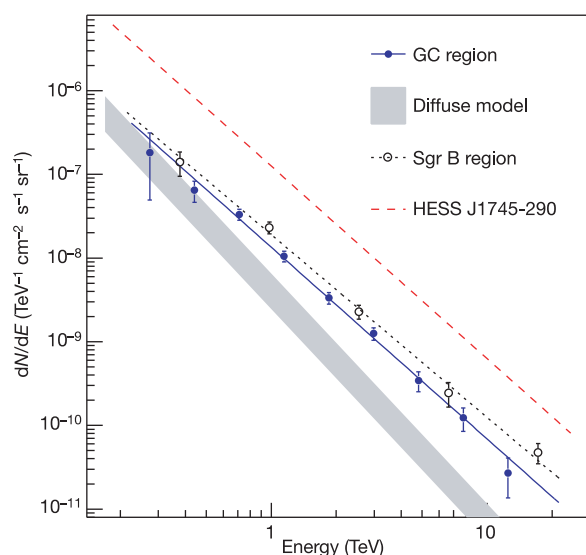


Figure 3 | Energy distribution of Galactic cosmic rays. γ -ray flux per unit angle in the Galactic Centre region (data points), compared with the expected flux, assuming a cosmic-ray spectrum as measured in the solar neighbourhood (shaded band). The spectrum of the region $|l| < 0.8^\circ$, $|b| < 0.3^\circ$ is shown using full circles. These data can be described by a power law: $dN/dE = k[E(\text{in TeV})]^{-\Gamma}$, with $k = (1.73 \pm 0.13_{\text{stat}} \pm 0.35_{\text{sys}}) \times 10^{-8} \text{ TeV}^{-1} \text{ cm}^{-2} \text{ s}^{-1} \text{ sr}^{-1}$ and a photon index $\Gamma = 2.29 \pm 0.07_{\text{stat}} \pm 0.02_{\text{sys}}$. The shaded box shows the range of expected π^0 -decay fluxes from this region assuming a cosmic-ray spectrum identical to that found in the solar neighbourhood and a total mass of $1.7\text{--}4.4 \times 10^7$ solar masses in the region $|l| < 0.8^\circ$, $|b| < 0.3^\circ$ estimated from CS measurements. Above 1 TeV an enhancement by a factor of 3–9 relative to this prediction is observed. Using independent mass estimates derived from submillimetre measurements²⁴, $5.3 \pm 1.0 \times 10^7$ solar masses, and from C^{18}O measurements²⁵, $3^{+2}_{-1} \times 10^7$ solar masses, results in enhancement factors of 4–6 and 5–13, respectively (see Supplementary Information). The strongest emission away from the bright central source HESS J1745–290 occurs close to the Sgr B complex of giant molecular clouds²⁶. In a box covering this region ($0.3^\circ < l < 0.8^\circ$, $-0.3^\circ < b < 0.2^\circ$), integrated CS emission suggests a molecular target mass of $6\text{--}15 \times 10^6$ solar masses. The energy spectrum of this region is shown using open circles. The measured γ -ray flux (>1 TeV) implies a high-energy cosmic-ray density which is 4–10 times higher than the local value. Standard γ -ray selection cuts are applied here, yielding a spectral analysis threshold of 170 GeV. The spectrum of the central source HESS J1745–290 is shown for comparison (using an integration radius of 0.14°). All error bars show ± 1 standard deviation.

for HESS) that would also be bright in the X-ray regime (as is, for example, G0.9+0.1). The existence of about ten such unknown sources in this small region again seems unlikely. Any substantially extended inverse Compton source would probably be a foreground source along the line of sight towards the Galactic Centre region, making any correlation with central molecular clouds entirely coincidental.

Received 13 July; accepted 22 November 2005.

- Ginzburg, V. L. (ed.) *Astrophysics of Cosmic Rays* (North Holland, Amsterdam, 1990).
- Hillas, A. M. Can diffusive shock acceleration in supernova remnants account for high-energy galactic cosmic rays? *J. Phys. G* **31**, R95–R131 (2005).
- Aharonian, F. Gamma rays from molecular clouds. *Space Sci. Rev.* **99**, 187–196 (2001).

- Hofmann, W. in *Proc. 28th ICRC, Tsukuba 2811–2814* (Univ. Academy Press, Tokyo, 2003).
- Hunter, S. D. et al. EGRET observations of the diffuse gamma-ray emission from the galactic plane. *Astrophys. J.* **481**, 205–240 (1997).
- Aharonian, F. A. et al. Very high energy gamma rays from the direction of Sagittarius A*. *Astron. Astrophys.* **425**, L13–L17 (2004).
- Tsuchiya, K. et al. Detection of sub-TeV gamma-rays from the Galactic Center direction by CANGAROO-II. *Astrophys. J.* **606**, L115–L118 (2004).
- Kosack, K. et al. TeV gamma-ray observations of the Galactic Center. *Astrophys. J.* **608**, L97–L100 (2004).
- Aharonian, F. A. et al. Very high energy gamma rays from the composite SNR G0.9+0.1. *Astron. Astrophys.* **423**, L25–L29 (2005).
- Aharonian, F. A. et al. The H.E.S.S. survey of the Inner Galaxy in very high energy γ -rays. *Astrophys. J.* (in the press) preprint at (<http://arXiv.org/astro-ph/0510397>) (2005).
- Tsuboi, M., Toshihiro, H. & Ukita, N. Dense molecular clouds in the Galactic Center region I. Observations and data. *Astrophys. J. Suppl.* **120**, 1–39 (1999).
- Oka, T. et al. A large-scale CO survey of the Galactic Center. *Astrophys. J. Suppl.* **118**, 455–515 (1998).
- Crocker, R. M. et al. The AGASA and SUGAR anisotropies and TeV gamma rays from the Galactic Center: A possible signature of extremely high energy neutrons. *Astrophys. J.* **622**, 892–909 (2005).
- Maeda, Y. et al. A Chandra study of the Sagittarius A east: A supernova remnant regulating the activity of our Galactic Center? *Astrophys. J.* **570**, 671–687 (2002).
- Rockefeller, G. et al. The X-ray ridge surrounding Sgr A* at the Galactic Center. *Astrophys. J.* (submitted); preprint at (<http://arXiv.org/astro-ph/0506244>) (2005).
- Aharonian, F. & Neronov, A. TeV gamma rays from the Galactic Center. *Space Sci. Rev.* (in the press); preprint at (<http://arXiv.org/astro-ph/0503354>) (2005).
- Béanger, G. et al. A persistent high-energy flux from the heart of the Milky Way: INTEGRAL's view of the Galactic Center. *Astrophys. J.* (in the press); preprint at (<http://arXiv.org/astro-ph/0508128>) (2005).
- Crutcher, R. M. Magnetic fields in molecular clouds: observations confront theory. *Astrophys. J.* **520**, 706–713 (1999).
- Morris, M. & Serabyn, E. The Galactic Center environment. *Annu. Rev. Astron. Astrophys.* **34**, 645–701 (1996).
- Mattox, J. R., Hartman, R. C. & Reimer, O. A quantitative evaluation of potential radio identifications for 3EG EGRET sources. *Astrophys. J. Suppl.* **135**, 155–175 (2001).
- Aharonian, F. A. et al. A new population of very high energy gamma-ray sources in the Milky Way. *Science* **307**, 1938–1942 (2005).
- Aharonian, F. A. et al. High-energy particle acceleration in the shell of a supernova remnant. *Nature* **432**, 75–77 (2004).
- Aharonian, F. A. et al. Detection of TeV gamma-ray emission from the shell-type supernova remnant RX J0852.0–4622 with HESS. *Astron. Astrophys.* **437**, L7–L10 (2005).
- Pierce-Price, D. et al. A deep submillimeter survey of the Galactic Center. *Astrophys. J.* **545**, L121–L125 (2000).
- Dahmen, G. et al. Molecular gas in the Galactic center region II. Gas mass and $\text{N}_2\text{H}^+/\text{I}_{12\text{CO}}$ conversion based on a $\text{C}^{18}\text{O}(J=1 \rightarrow 0)$ survey. *Astron. Astrophys.* **331**, 959–976 (1998).
- Lis, D. C. & Goldsmith, P. F. CO isotope studies and mass of the Sagittarius B2 molecular cloud. *Astrophys. J.* **337**, 704–711 (1989).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The support of the Namibian authorities and of the University of Namibia in facilitating the construction and operation of HESS is gratefully acknowledged, as is the support by the German Ministry for Education and Research (BMBF), the Max Planck Society, the French Ministry for Research, the CNRS-IN2P3 and the Astroparticle Interdisciplinary Programme of the CNRS, the UK Particle Physics and Astronomy Research Council (PPARC), the IPNP of Charles University, the South African Department of Science and Technology and National Research Foundation, and by the University of Namibia. We would like to thank M. Tsuboi for providing the CS survey data used here and Y. Moriguchi and Y. Fukui for helpful discussions on molecular tracers.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.A.H. (jim.hinton@mpi-hd.mpg.de).

Anti-planetward auroral electron beams at Saturn

J. Saur^{1†}, B. H. Mauk¹, D. G. Mitchell¹, N. Krupp², K. K. Khurana³, S. Livi¹, S. M. Krimigis¹, P. T. Newell¹, D. J. Williams¹, P. C. Brandt¹, A. Lagg², E. Roussos² & M. K. Dougherty⁴

Strong discrete aurorae on Earth are excited by electrons, which are accelerated along magnetic field lines towards the planet¹. Surprisingly, electrons accelerated in the opposite direction have been recently observed^{2–6}. The mechanisms and significance of this anti-earthward acceleration are highly uncertain because only earthward acceleration was traditionally considered, and observations remain limited. It is also unclear whether upward acceleration of the electrons is a necessary part of the auroral process or simply a special feature of Earth's complex space environment. Here we report anti-planetward acceleration of electron beams in Saturn's magnetosphere along field lines that statistically map into regions of aurora. The energy spectrum of these beams is qualitatively similar to the ones observed at Earth, and the energy fluxes in the observed beams are comparable with the energies required to excite Saturn's aurora. These beams, along with the observations at Earth^{2–6} and the barely understood electron beams in Jupiter's magnetosphere^{7,8}, demonstrate that anti-planetward acceleration is a universal feature of aurorae. The energy contained in the beams shows that upward acceleration is an essential part of the overall auroral process.

Measurements by the Cassini spacecraft^{9,10} in Saturn's magnetosphere reveal electron distributions on near equatorial segments that contain electron beams (Fig. 1). These electron beams are strongly aligned with respect to the local magnetic field and have great similarity to those observed at Earth^{2–6}.

In order to constrain the spatial occurrence of the electron beams, we examine all available data of the first four Cassini spacecraft orbits around Saturn (identified as SOI (Saturn orbit insertion), 0a, 0b, and 0c). The data are binned into different categories, which we display as a function of their position in Saturn's magnetosphere (Fig. 2). Segments of Cassini's orbit marked red display regions where we find magnetic-field-aligned electron beams. On the inbound pass—that is, the dawn–noon sector of the magnetosphere—we see on two orbits (0a, 0b) electron beams at radial distances between 11 R_S and 23 R_S (R_S indicates Saturn's radius). Measurements during the other two orbits do not satisfy our observational constraints and render the electron distribution undetermined. On the outbound pass—that is, in the post-midnight to dawn sector—we find electron beams during orbits SOI, 0a and 0c at radial distances between 19 R_S and 33 R_S . The spacecraft attitude on other parts of the outbound orbits does not seriously constrain the electron distribution. On orbital segments with electron beams we do not simultaneously observe ion beams.

The electron beaming is observed over all available energies from 20 keV to 800 keV, and not just over a narrow range of energy (Fig. 3). The energy distributions of the beams follow roughly a power law as a function of energy, with spectral index between -1.6 and -2.6 . The broad nature of the energy spectrum suggests a stochastic acceleration process and not a simple potential drop along the magnetic field

lines, consistent with the analogous Earth beams^{2–6}. While the spacecraft is moving radially inward during orbit 0a, the energy flux of the beams increases by about a factor of 10 at the lowest energies.

Because of the similarity of the observed beams to the beaming observed at Earth, both in terms of the narrowness of the beaming

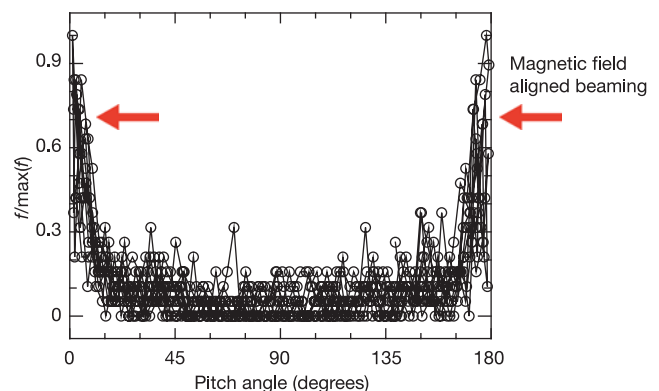


Figure 1 | Segments of Cassini's orbit in Saturn's magnetosphere show electron distributions that contain magnetic-field-aligned electron beams. This distribution was taken during a 10 min segment of Cassini's inbound orbit 0a at 11:27 on 26 October 2004 at a radial distance of 21.3 R_S with a priority channel in an energy range from 28.1 to 49 keV. Zero and 180° are along the locally measured magnetic field²⁸, which is where we find electron beams. Electron beam intensities f are displayed normalized to the maximum value $\max(f)$ of the segment. The electron data are measured by the low energy telescope of the Low Energy Magnetospheric Measurement System (LEMMS)^{9,10} on board the Cassini spacecraft. It measures the electron distribution relative to the magnetic field direction in different energy channels within a total energy range from 20.2 keV to 829 keV and an instrument full-width cone of 15°. A detailed analysis of the angular width of the equatorial beams, including the detector response function, provides an upper limit of 10° for the angular half width, but permits any smaller angle. A source distribution close to Saturn that is narrow in angular width at its source would be consistent with our observed maximum 10° width. This still holds if pitch angle scattering is included. With a pitch angle diffusion coefficient²⁹ $D_0 = 5 \times 10^{-3} \text{ s}^{-1}$ taken from Jupiter's magnetosphere at a comparable radial distance, we estimate that beams with an energy of ~ 30 keV widen by 8° while travelling from Saturn out to the location of the Cassini spacecraft. If we assume alternatively that the source electron distribution is not only broad in energy but also in its angular distribution, that is, isotropic, then the action of the magnetic mirror force requires the electrons to be accelerated close to Saturn in order to be field-aligned at the equator. Under the assumption of isotropy, a maximum 10° width at the spacecraft places the electron acceleration region somewhere below 5 R_S from Saturn's atmosphere. Additional pitch angle scattering requires the origin of the beams to be much closer to Saturn.

¹Applied Physics Laboratory, Johns Hopkins University, 11100 Johns Hopkins Road, Laurel, Maryland 20723, USA. ²Max-Planck-Institut für Sonnensystemforschung, 37191 Katlenburg-Lindau, Germany. ³Institute of Geophysics and Planetary Physics, UCLA, Los Angeles, California 90095, USA. ⁴Blackett Laboratory, Imperial College, London SW7 2AZ, UK. [†]Present address: Institut für Geophysik und Meteorologie, Universität zu Köln, Albertus-Magnus-Platz, 50923 Köln, Germany.

and the breadth of the energies involved, our most reasonable inference is that the origin of the electron beams is close to Saturn, in analogy with the demonstrated origin of such beams at Earth^{2–6} (see also Fig. 1 legend). Similar conclusions have been reached for the origin of electron beams in Jupiter's magnetosphere⁷ and in the wake of Jupiter's satellite Io¹¹. Independent support for a near Saturn source comes from Cassini radio observations^{12,13}. These place the source region of Saturn's kilometric radiation, which is associated with Saturn's auroral electron acceleration¹², over its auroral zone.

In order to investigate the longitudes and latitudes where the beams originate near Saturn, we use the latest magnetic field model¹⁴ developed with Cassini and Voyager data. In Fig. 4, we show the electron beams, measured near the equator, mapped with the magnetic field model into Saturn's atmosphere as red data points. We overlay the projected locations of the beams with two auroral images¹⁵, obtained before the Cassini period when the beams were observed. No auroral images are available for the same time period when the Cassini spacecraft made the electron measurements. Although Saturn's aurorae vary in time^{16,17}, our auroral images (Fig. 4) as well as 68 recent further images¹⁸ provide a good statistical framework for the average location of Saturn's auroral features. Auroral activity occurs over a range of latitudes that correspond well to the spatial range where we find electron beams, that is, from the magnetopause to as far inside as $11R_S$. In these images it appears that main features of Saturn's aurorae are located at latitudes between 70° and 75° , which correspond to Saturn's middle magnetosphere, according to our field model. For example, the red data points near 80° latitude on the noon side (Fig. 4) characterize electron beams just inside the magnetopause boundary. Saturn's aurorae lie in general

well equatorward of this boundary. From the recent 68 images¹⁸, an extended reference oval for quiet solar wind conditions¹⁹ at 72° could be extracted, which corresponds to $12R_S$ in the noon sector of the magnetosphere, where we observe electron beams. This shows that main features of Saturn's aurorae originate inside the magnetosphere significantly planetward of its magnetopause. Since the mapping of the beams depends on the magnetic field model, we independently also use a simple dipole magnetic field. The results are shown with yellow data points in Fig. 4. With this simple model, some regions giving rise to Saturn's aurorae need to lie even further inside $11R_S$. Beams at a radial distance of $11R_S$, however, appear to be too far away from the open/closed field line boundaries or the tail of the magnetosphere to be created there, the long-standing expectation for the source regions of Saturn's aurorae^{20–22}. Our findings suggest that all three planets—Saturn, Jupiter²³ and Earth^{24,25}—have (to a varying degree) contributions to their auroral source regions well inside their magnetospheric boundaries.

For the observed near-equatorial energy spectra of Fig. 3, we estimate the corresponding energy flux at the low altitude source region, that is, close to Saturn (Fig. 3 legend). We do not measure electrons with energies less than 20 keV, but owing to the beams' power law behaviour, we expect a significant flux below 20 keV. If we therefore extrapolate the spectra to 2 keV, we find an energy flux of 0.13 mW m^{-2} close to Saturn, based on measurements at $21.4R_S$ (0.1 mW m^{-2} not extrapolated). The projected fluxes at Saturn increase strongly for measurements taken closer to Saturn. For example, we find 1.5 mW m^{-2} at $13.7R_S$ also extrapolated to 2 keV (0.3 mW m^{-2} not extrapolated). If we assume that the electron beam intensity correlates with auroral intensity, then the higher energy fluxes further inside of Saturn's magnetosphere additionally support the idea that main features of Saturn's aurorae originate deep inside its magnetosphere.

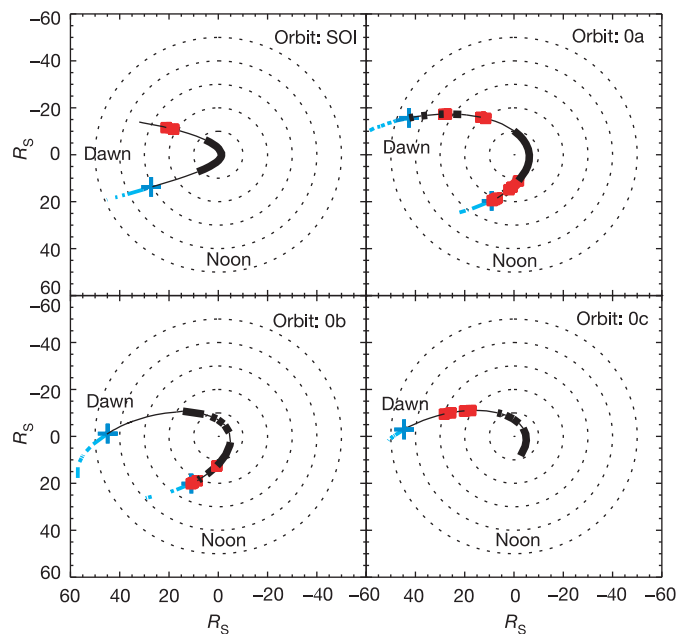


Figure 2 | Electron beams in Saturn's magnetosphere. Locations of electron beams in Saturn's magnetosphere for the first four orbits of Cassini's tour are shown in red. Segments in solid black indicate locations that we positively identify not to be field-aligned; segments in thin black, undetermined distribution (mostly owing to insufficient data coverage); dark blue cross, innermost magnetopause crossing; turquoise, magnetosheath; dotted circles are separated by $10R_S$ (R_S , Saturn's radius) and are for orientation purposes only. The x-axis is perpendicular to Saturn's spin axis and points in direction to the Sun. The y-axis is chosen so that the x-axis, the y-axis and Saturn's spin axis display a right-handed coordinate system. Cassini travels in an anticlockwise fashion in these panels.

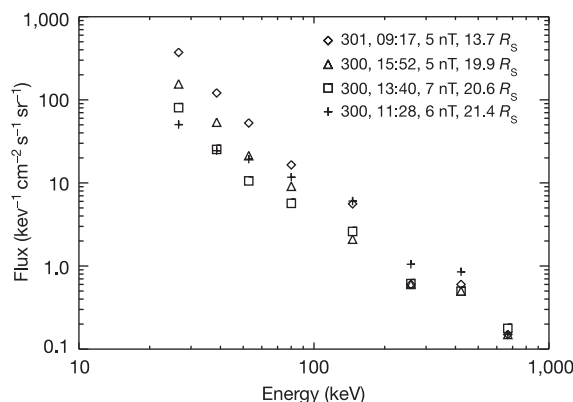


Figure 3 | Four electron beam energy spectra during Cassini's inbound orbit 0a in 2004. Each spectrum is calculated from the electron distributions (similar to Fig. 1) of eight different energy channels within the instrument energy range of 20–800 keV; four parameters for each spectrum are given in the key, namely, the day of 2004, the time in the day, the magnetic field and the radial distance from Saturn. In each distribution we only calculate the energy within the beam, that is, we only count the electron flux that lies within an 11.25° cone with respect to the local magnetic field. We can use these spectra to estimate the total electron energy fluxes at the low altitude source region (close to Saturn) by taking into account the characteristics of the LEMMS detector and the width of the loss cone. We assume that the total energy flux within a magnetic flux tube is equal at the equator and at the origin of the acceleration. The total low altitude energy flux (20–800 keV) for the spectrum at 13.7, 19.9, 20.6 and $21.4 R_S$ is respectively 0.34, 0.16, 0.1 and 0.1 mW m^{-2} ($= 10^{-3} \text{ W m}^{-2}$), with a spectral power law slope of respectively -2.6 , -2.5 , -2.1 and -1.6 . We expect the energy spectrum to extend (as a similar power law) to lower energies. When we extrapolate the spectra to 2 keV, we find a total electron energy flux of 1.5, 0.6, 0.18 and 0.13 mW m^{-2} , respectively.

Saturn's observed auroral ultraviolet intensities are typically in the range of 1 kilorayleigh (1 kR) to up to 70 kR (ref. 16), which corresponds to an energy flux of $0.1\text{--}7\text{ mW m}^{-2}$. The energy fluxes that we infer from the observed electron beams of $0.1\text{--}1.5\text{ mW m}^{-2}$ are surprisingly large, and comparable to the energies needed to excite Saturn's aurorae.

We conclude that electron beams are accelerated away from Saturn on field lines that map into the auroral regions of Saturn. At Earth, anti-planetward accelerated electrons have also been observed^{2–6}.

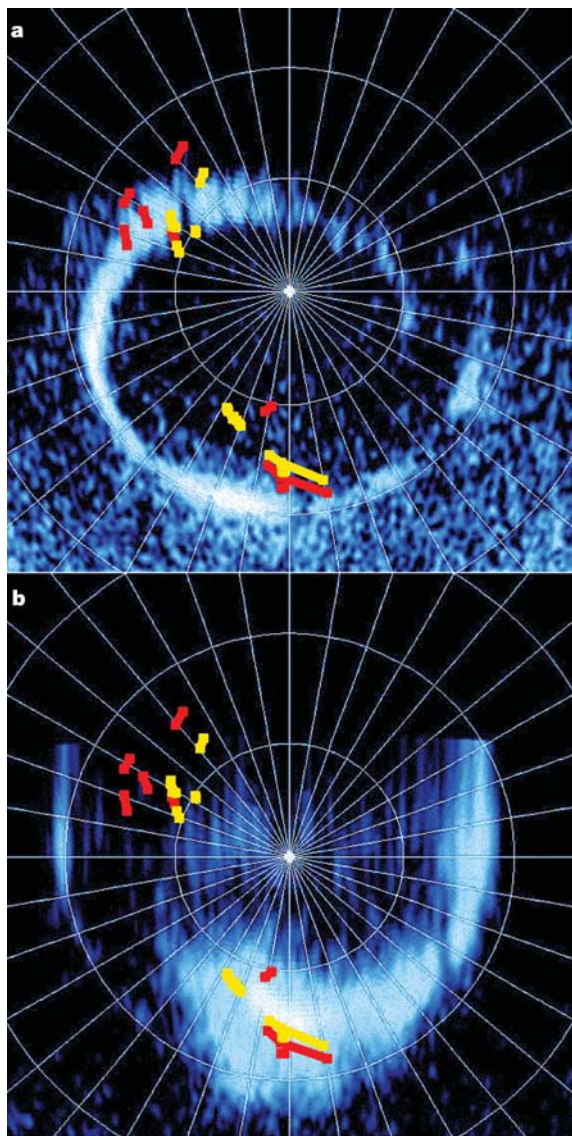


Figure 4 | Saturn's electron beams and aurorae. Shown are the locations of all observed electron beams that mapped into Saturn's polar atmosphere (red and yellow data points) over-plotted in comparison with two snapshots¹⁵ of Saturn's aurora, **a** and **b**. Data points in red are calculated with the latest magnetic field model¹⁴, data points in yellow are calculated with a simple dipole magnetic field. To the left is the morning side of Saturn and bottom is towards noon; circles display 70° and 80° planetary latitude. Saturn's aurora varies in time, but electron beams map in a statistical sense well into regions of Saturn's aurora. As a technical note, in the Khurana magnetic field model¹⁴ the dayside magnetopause is located at $20R_S$. For orbits 0a and 0b, the magnetopause is located at $21.6R_S$ and $22.9R_S$. Therefore a few locations of electron beams cannot be mapped with this magnetic field model into Saturn's atmosphere, and are neglected in this figure. (This figure has been reproduced from figure 3 of ref. 15. Copyright 2004 American Geophysical Union. Reproduced/modified by permission of the AGU.)

Their significance and their mechanisms remain however, uncertain, both owing to a lack of comprehensive studies and owing to the fact that currently no generally accepted model for the upward acceleration exists, although a few ideas have been proposed^{26,27}. With the observations at Earth only, it is not clear whether anti-planetward accelerated auroral electrons are unique to the special properties of the complex Earth environment or if they are a general property of aurorae. Recently, electron beams have also been observed in Jupiter's magnetosphere^{7,8}. Their origin and their physical relation to Jupiter's aurorae are, however, unclear, and the huge qualitative differences between Jupiter's and Earth's magnetosphere bring into question the extent to which the beams at Earth and Jupiter are analogous. Observations of beams at Saturn, that is, in a magnetospheric configuration that lies between that of Earth and Jupiter, suggest that anti-planetward acceleration of electrons is a universal property of aurorae. The energy fluxes in the anti-planetward accelerated beams compete with the fluxes contained in the beams that are directed towards the planet, and therefore generate aurorae. Thus in addition to the universality of anti-planetward acceleration, Saturn's aurorae suggest that anti-planetward acceleration is also a key player in the energetics of the auroral process.

Received 13 June; accepted 2 November 2005.

- Paschmann, G., Haaland, S. & Treumann, R. *Auroral Plasma Physics* (Kluwer Academic, Dordrecht, 2003).
- Klumpar, D. M. in *Physics of Space Plasmas* (1989) (eds Chang, T., Crew, G. B. & Jasperse, J. R.) 265–276 (SPI Conference Proceedings and Reprint Series No. 9, Scientific Publishers, Cambridge, Massachusetts, 1990).
- Klumpar, D., Quinn, J. & Shelley, E. Counter-streaming electrons at the geomagnetic equator near 9 earth radii. *Geophys. Res. Lett.* **15**, 1295–1298 (1988).
- Carlson, C. et al. FAST observations in the downward auroral current region: Energetic upgoing electron beams, parallel potential drops and ion heating. *Geophys. Res. Lett.* **25**, 2017–2020 (1998).
- Ergun, R. et al. FAST satellite observations of electric field structures in the auroral zone. *Geophys. Res. Lett.* **25**, 2025–2028 (1998).
- Marklund, M. et al. Temporal evolution of the electric field accelerating electrons away from the auroral ionosphere. *Nature* **414**, 724–727 (2001).
- Frank, L. A. & Paterson, W. R. Galileo observations of electron beams and thermal ions in Jupiter's magnetosphere and their relationship to the auroras. *J. Geophys. Res.* **107**, A1478, doi:10.1029/2001JA009150 (2002).
- Tomás, A. et al. Energetic electrons in the inner part of the Jovian magnetosphere and their relation to auroral emissions. *J. Geophys. Res.* **109**, A203, doi:10.1029/2004JA010405 (2004).
- Krimigis, S. et al. Magnetosphere Imaging Instrument (MIMI) on the Cassini Mission to Saturn/Titan. *Space Sci. Rev.* **114**, 223–329 (2004).
- Krimigis, S. et al. Dynamics of Saturn's magnetosphere from MIMI during Cassini's orbital insertion. *Science* **307**, 1270–1273 (2005).
- Williams, D. J., Thorne, R. M. & Mauk, B. H. Energetic electron beams and trapped electrons at Io. *J. Geophys. Res.* **104**, 14739–14753 (1999).
- Kurth, W. S. An Earth-like correspondence between Saturn's auroral features and radio emission. *Nature* **443**, 722–725 (2005).
- Gurnett, D. A. et al. Radio and plasma wave observations at Saturn from Cassini's approach and first orbit. *Science* **307**, 1255–1259 (2005).
- Khurana, K. K., Arridge, C. S. & Dougherty, M. K. A versatile model of Saturn's magnetospheric field. *Geophys. Res. Abstr.* **7**, 05970 (2005); (<http://www.cosis.net/abstracts/EGU05/05970/EGU05-J-05970.pdf>).
- Gérard, J.-C. et al. Characteristics of Saturn's FUV aurora observed with the Space Telescope Imaging Spectrograph. *J. Geophys. Res.* **109**(A9), A09207, doi:10.1029/2004JA010513 (2004).
- Clarke, J. T. et al. Morphological differences between Saturn's ultraviolet aurorae and those of Earth and Jupiter. *Nature* **433**, 717–719 (2005).
- Prangé, R. et al. An interplanetary shock traced by planetary auroral storms from the Sun to Saturn. *Nature* **432**, 78–81 (2004).
- Grodent, D., Gérard, J.-C., Cowley, S. W. T., Bunce, E. J. & Clarke, J. T. Variable morphology of Saturn's southern ultraviolet aurora. *J. Geophys. Res.* **110**, A07215, doi:10.1029/2004JA010983 (2005).
- Crary, F. J. et al. Solar wind dynamic pressure and electric field as the main factors controlling Saturn's aurorae. *Nature* **433**, 720–722 (2005).
- Bhardwaj, A. & Gladstone, G. R. Auroral emissions of the giant planets. *Rev. Geophys.* **38**, 295–353 (2000).
- Cowley, S. T. W., Bunce, E. & Prangé, R. Saturn's polar ionospheric flows and their relation to the main auroral oval. *Ann. Geophys.* **22**, 1379–1394 (2004).
- Cowley, S. T. W. et al. Reconnection in a rotation-dominated magnetosphere and its relation to Saturn's auroral dynamics. *J. Geophys. Res.* **110**, A00201, doi:10.1029/2004JA010796 (2005).

23. Clarke, J. T. *et al.* Ultraviolet emissions from the magnetic footprints of Io, Ganymede and Europa on Jupiter. *Nature* **415**, 997–1000 (2002).
24. Feldstein, Y. I. & Galperin, Y. I. The auroral luminosity in the high-latitude upper atmosphere: its dynamics and relationship to the large-scale structure of the Earth's magnetosphere. *Rev. Geophys.* **23**, 217–275 (1985).
25. Mauk, B. H. & Meng, C.-I. in *Auroral Physics* (eds Meng, C. I., Rycroft, M. J. & Frank, L. A.) 223–239 (Cambridge Univ. Press, Cambridge, 1991).
26. Temerin, M. & Carlson, C. W. Current-voltage relationship in the downward auroral current region. *Geophys. Res. Lett.* **25**, 2365–2368 (1998).
27. Anderson, L. *et al.* Characteristics of parallel electric fields in the downward current region of the aurora. *Phys. Plasmas* **9**, 3600–3609 (2002).
28. Dougherty, M. K. *et al.* The Cassini Magnetic Field Investigation. *Space Sci. Rev.* **114**, 331–383 (2004).
29. Williams, D. J. & Mauk, B. H. Pitch angle diffusion at Jupiter's moon Ganymede. *J. Geophys. Res.* **102**, 24283–24287 (1997).

Acknowledgements We thank D. Grodent for helpful discussions.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.S. (saur@geo.uni-koeln.de).

Electronic transport in nanometre-scale silicon-on-insulator membranes

Pengpeng Zhang¹, Emma Tevaarwerk¹, Byoung-Nam Park¹, Donald E. Savage¹, George K. Celler², Irena Knezevic¹, Paul G. Evans¹, Mark A. Eriksson¹ & Max G. Lagally¹

The widely used 'silicon-on-insulator' (SOI) system consists of a layer of single-crystalline silicon supported on a silicon dioxide substrate. When this silicon layer (the template layer) is very thin, the assumption that an effectively infinite number of atoms contributes to its physical properties no longer applies, and new electronic, mechanical and thermodynamic phenomena arise^{1–4}, distinct from those of bulk silicon. The development of unusual electronic properties with decreasing layer thickness is particularly important for silicon microelectronic devices, in which (001)-oriented SOI is often used^{5–7}. Here we show—using scanning tunnelling microscopy, electronic transport measurements, and theory—that electronic conduction in thin SOI(001) is determined not by bulk dopants but by the interaction of surface or interface electronic energy levels with the 'bulk' band structure of the thin silicon template layer. This interaction enables high-mobility carrier conduction in nanometre-scale SOI; conduction in even the thinnest membranes or layers of Si(001) is therefore possible, independent of any considerations of bulk doping, provided that the proper surface or interface states are available to enable the thermal excitation of 'bulk' carriers in the silicon layer.

The use of SOI in microelectronics and nanoelectromechanical systems is already pervasive. SOI promises, in fact, to become the platform for future high-speed electronics as well as for a range of sensor technologies^{5,7}. Many of the expected advances will depend on making the Si template layer in SOI increasingly thin. In terms of electronic-transport properties, such very thin crystalline layers pose unique challenges. When the Si layer becomes thin enough, charge traps at the oxide/Si interface will deplete the Si layer of free carriers, making the resistivity high. Even for very-high-quality oxide/Si interfaces, the density of these charge traps is of the order of 10^{10} – 10^{11} cm^{−2}. In contrast, a 10 nm Si layer that is doped at a typical 10^{15} cm^{−3} has only 10^9 dopant atoms per cm². One would therefore predict that measurements depending on the flow of current (such as scanning tunnelling microscopy (STM), electric-force microscopy and Hall effect) and electron emission and reflection processes (such as photoelectron spectroscopy and low-energy electron microscopy) become impossible. Previous reports have, in fact, claimed that surface and interface charge trapping states make STM impossible on thin SOI^{8,9}. We show that STM imaging of what is conventionally considered as fully depleted SOI is indeed possible on clean SOI. We demonstrate that electronic conduction is enabled by reconstruction of the clean Si(001) surface, which shifts the Fermi level and results in a population of charge carriers. We suggest that the phenomenon is general if appropriate states are available.

Clean SOI consists of a thin Si layer bounded by two interfaces: Si/SiO₂ and Si/vacuum. The Si/SiO₂ interface quality is determined by the SOI fabrication process; it is very good indeed in bonded SOI¹⁰. We used SOI made both by wafer bonding and by the SIMOX process

(implant with oxygen and anneal to form a buried oxide). The template layer was doped with boron nominally at 10^{15} cm^{−3}, and was thinned by dry thermal oxidation at 1,050 °C, followed by wet chemical etching to remove the oxide. To produce thicknesses below 20 nm we used repeated oxidation and oxide etching steps. Very-low-defect-density clean surfaces were produced by removing the final protective oxide via the slow deposition of several monolayers (ML) of Si or Ge at 700 °C in ultrahigh vacuum followed by thermal annealing (see Methods).

Figure 1 shows a schematic diagram of the STM process and images of the surface of several ultrathin-template SOI samples prepared in the above manner. The images shown here were acquired at room temperature in the constant-current mode using a tunnelling current of 0.1 nA. Filled-state and empty-state images of the Si surface were obtained at tip biases of −2 V and +2 V, respectively, relative to the sample. The images are comparable to those obtained from very clean bulk Si(001) at the same bias voltage¹¹. The surface reconstructs to form rows of dimerized atoms¹² with alternate orthogonal (2 × 1) and (1 × 2) terraces. When we clean the surface with Ge, a zigzag pattern indicative of dimer buckling appears (Fig. 1c) in both filled- and empty-state images, just as in bulk Si(001) on which a small amount of Ge has been deposited¹³. The comparatively rougher Si/SiO₂ interface in SIMOX wafers does not affect the quality of STM images of the surfaces of SIMOX template layers (Fig. 1d). Strained SOI also produces outstanding STM images (Fig. 1e).

We wish to use STM as a probe of the conductivity of the thin Si template layer. In order to do so, we must show to what degree this approach is sensible. The two important steps in the conduction of electrons during STM imaging (Fig. 1a) are (1) the tunnelling of an electron from tip to sample (or vice versa) and (2) the subsequent removal of the tunnelling charge to a distant electrical lead to complete the circuit. In the complete absence of the second step, STM is impossible. If a very large sample resistance impedes the second step, imaging will require higher applied voltages. The filled- and empty-state images in Fig. 1 show that on clean SOI STM is not only possible, but requires no excess voltage: the images are acquired with the same minimum voltage used for imaging bulk Si. The effective resistance of the Si layer in STM imaging of thin SOI is thus much less than the tunnelling resistance (20 GΩ) in our STM measurements.

To understand the role a Si/SiO₂ interface plays in the Si layer's electronic properties, we analysed the sheet resistance of Si membranes with thicknesses ranging from 15 nm to 200 nm, sandwiched between a native oxide and the buried oxide. The nominal doping level of the Si layer is 10^{15} cm^{−3}. The interface trap density of states (D_{it}) typically exhibits a 'U' shape across the Si bandgap, with D_{it} of the order of 10^{10} to 10^{11} cm^{−2} eV^{−1}. Interface traps in the upper half of the bandgap behave as acceptors, while those in the lower half

¹University of Wisconsin-Madison, Madison, Wisconsin 53706, USA. ²Soitec USA, 2 Centennial Drive, Peabody, Massachusetts 01960, USA.

behave as donors^{14,15}. A thin Si layer is depleted of free carriers, which results in a high sheet resistance. Using the van der Pauw method, we have measured a two-dimensional (2D) resistivity $\rho_{2D} = 80 \text{ G}\Omega\text{square}^{-1}$ for a 20 nm Si layer sandwiched between a native oxide and the buried oxide, a reasonable value for such thin SOI¹⁶.

The STM imaging occurs, of course, with the top oxide removed. The simplest explanation for the good STM images might be that removal of the top (native) oxide reduces the number of traps by a factor of two, freeing some of the trapped carriers. The Si/SiO₂ interface has two orders of magnitude more traps than needed to deplete our samples, however, and therefore a factor-of-two reduction in trap density has an insignificant direct effect on carrier populations. In our STM experiments, the sample electrical lead is far from the tunnelling tip, and the contribution of the sample resistance to the STM circuit is, to a good approximation, the resistance R between two concentric circles, one with the radius at which the current from the tip reaches the sample, r_1 , and the other at the radius of the contacts, r_2 , giving:

$$R = \int_{r_1}^{r_2} \rho_{2D} \frac{dr}{2\pi r} = \frac{\rho_{2D}}{2\pi} \ln\left(\frac{r_2}{r_1}\right) \quad (1)$$

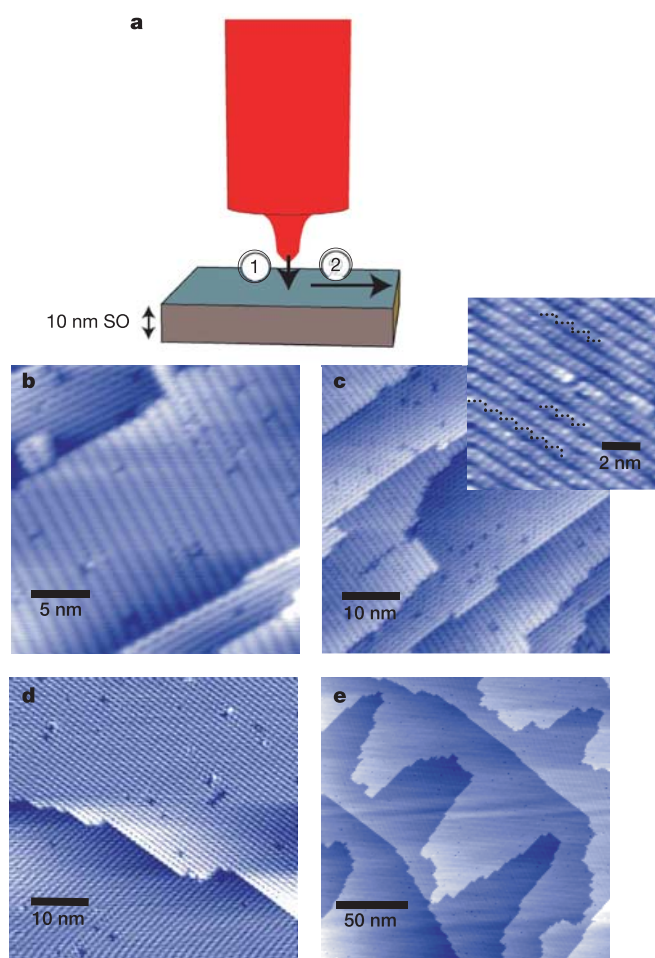


Figure 1 | Scanning tunnelling microscopy experiment and images of surfaces of clean SOI. **a**, Schematic diagram of the circuit completed by the tunnelling-gap (1) and sample (2) resistances. **b**, Filled-state STM image of a 10-nm-thick Si template layer in bonded SOI(001) with native oxide removed by ~3 ML of Si (see 'Methods'). **c**, Same as **b** but using Ge to remove the oxide. Inset, empty-state image showing a dimer buckling zigzag pattern produced by a fractional monolayer of Ge remaining in the Si(001) surface. **d**, Filled-state image of the surface of 20-nm-thick SIMOX SOI(001). **e**, Filled-state image of the surface of 15-nm-thick bonded strained SOI(001) (0.8% tensile strain).

The value of R depends only logarithmically on r_2 and r_1 , and therefore details such as the sample shape and the current flow from the tip into the sample are unimportant. Taking reasonable values $r_1 = 1 \text{ nm}$ and $r_2 = 0.5 \text{ cm}$, we find that $R \approx 2.5\rho_{2D} = 200 \text{ G}\Omega$. This value is far too large, in comparison to the total resistance of the STM circuit, to enable our STM measurements.

An additional conduction mechanism must therefore exist to enable STM on the clean Si(001) template surface. STM imaging is possible because of the influence of the surface π bands formed by the dimer bond reconstruction¹⁷. Charge transfer inside the tilted dimers on the surface results in an almost filled π band and an almost empty π^* band, each with the density of states of about $10^{15} \text{ cm}^{-2} \text{ eV}^{-1}$. The gap of about 0.5 eV between these two bands is positioned close to the valence band. This surface gap and the bottom of the π^* band pull the Fermi level downward, leading to a large concentration of holes in the Si template layer valence band via thermal excitation of electrons to the π^* band. The electronic band structure is shown schematically in Fig. 2 for Si layers with a native oxide and with a clean surface. Conduction in the surface states is immaterial to the Si layer conduction. The conduction of charge carriers by surface states has been probed with a variety of experimental techniques^{18–20}. Unless the mobility of electrons in surface bands is large enough to be comparable to the extremely high mobility of holes and electrons in Si, the contribution of surface state conduction to the electronic-transport properties of thin SOI will be negligible.

The combination of the conduction due to holes created in the valence band by excitation of charges into the close-lying empty surface states and the (possibly much lower) conduction of electrons in the surface π^* band results in an effective resistance R_{eff} (refs 21, 22):

$$\frac{1}{R_{\text{eff}}} = \frac{1}{R'_{\text{bulk}}} + \frac{1}{R_{\text{surface}}} \quad (2)$$

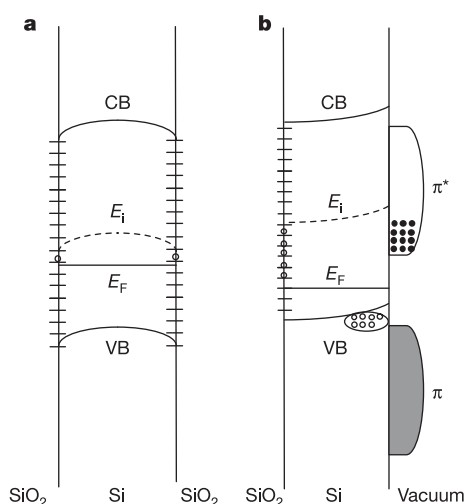


Figure 2 | Proposed band diagrams showing interface, bulk and surface bands for ultrathin SOI. **a**, Ultrathin SOI bounded by two Si/SiO₂ interfaces. For a 10 nm Si film, the maximum band bending is $<100 \mu\text{eV}$ (grossly exaggerated in the figure). The Fermi level is ~25 meV below the intrinsic level. **b**, Ultrathin SOI bounded by the reconstructed Si(001) surface on one face. The existence of the surface bands results in a dramatically reduced effective bandgap (~0.3 eV) between the (bulk) valence band of the thin Si layer and the surface π^* band. The enhanced conductivity is derived from the corresponding thermal population of carriers in those two bands. For a 10 nm Si film, the maximum band bending is ~7 meV. The Fermi level is ~0.4 eV below the intrinsic level at the Si/SiO₂ interface. CB, conduction band; VB, valence band; E_F , Fermi energy (Fermi level); E_i , energy of centre of bandgap (intrinsic level). Short lines denote interface trap states. Holes are indicated by open circles, and in **b** electrons are indicated by filled circles.

where $1/R'_{\text{bulk}}$ is the conductance arising from the thermal population of holes created by repositioning the Fermi level, and $1/R_{\text{surface}}$ is the conductance arising from the thermal population of electrons in the surface π^* states. Carriers at the buried Si/SiO₂ interface are effectively immobile.

The value of R_{eff} is much less than the 200 G Ω found for oxidized surfaces. For the range of acceptable values of the density of interface traps, D_{it} , from Fermi distribution calculations, the hole concentration in the valence band of the 10 nm SOI layer is $(2\text{--}4) \times 10^{10} \text{ cm}^{-2}$, the electron density in the π^* band is $(4.4\text{--}6) \times 10^{10} \text{ cm}^{-2}$, and the net remaining charge is in the form of trapped holes at the Si/SiO₂ interface. The overall distribution of charges is summarized in Fig. 2b. Charge carriers in the Si valence band and at the surface are able to respond to applied electric fields, and the calculated carrier concentrations are a basis for a quantitative comparison of R'_{bulk} and R_{surface} . The high-mobility holes are not influenced by the orientation of the local surface reconstruction.

Figure 3 shows the variation of the sheet resistance as a function of Si(001) template layer thickness in SOI. The top curve is a fit to the measured resistivity of the Si(001) layer bounded by oxide on both sides. The bottom curve is a calculation of the sheet resistance for a Si(001) layer with a clean reconstructed surface. The predictions of Fig. 3 are based on a numerical estimate of the band bending, assuming a surface density of states given by first-principles calculations of the Si(001) surface electronic structure¹⁷. We have assumed a density of states at the Si/SiO₂ interface that is consistent with previous measurements¹⁰. When the high mobility of carriers introduced into the Si 'bulk' by the surface bands is considered, the resistivity is considerably lower than the expected resistivity due to dopants alone for all Si film thicknesses below 70 nm. For both the

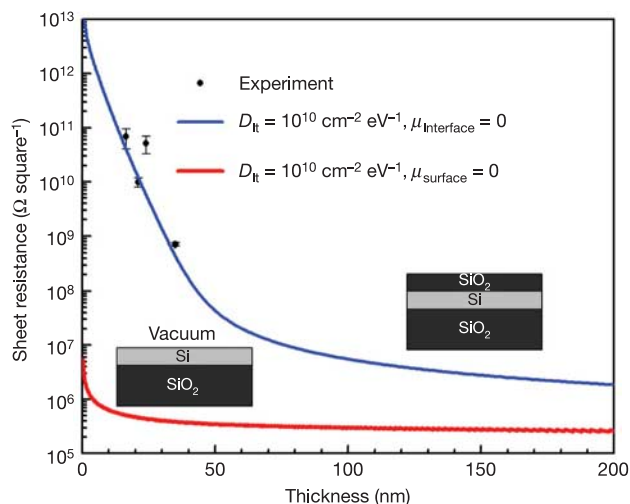


Figure 3 | Sheet resistance, as function of the Si layer thickness, of a thin Si membrane with different bounding layers. Upper curve, bounded by two Si/SiO₂ interfaces (right inset structure diagram); lower curve, bounded on one side by a clean reconstructed Si(001) surface (left inset structure diagram). Error bars ($\pm$ s.d.) give the uncertainty in the measured sheet resistance values—they are higher for higher resistances. The upper curve is a fit to the resistance measurements, requiring a density of local interface trap states $D_{\text{it}} = 1.1 \times 10^{10} \text{ cm}^{-2} \text{ eV}^{-1}$, a value that corresponds to high-quality Si/SiO₂ interfaces and an interface state mobility, $\mu_{\text{interface}}$, of zero. The lower curve is calculated with the same D_{it} and assuming that the mobility of carriers in the surface states, μ_{surface} , is negligible. The sheet resistance is not sensitive to the value of the surface state mobility: even a very high surface state mobility has an insignificant effect on the total conductivity. For very thin Si membranes, therefore, the existence of appropriate surface bands leads to a conductivity dramatically higher than expected for a thin layer of Si with two oxide terminations. A difference in sheet resistance of at least an order of magnitude persists to large Si membrane thicknesses.

oxide- and the vacuum-terminated ultrathin SOI, the bulk doping density is virtually irrelevant for electronic properties.

It is thus the interaction between surface and bulk that enables high-mobility conduction in nanoscale SOI. This interaction suggests that conduction in even the thinnest membranes or layers of Si(001) is possible, independent of any considerations of bulk band structure, defect states, and doping, as long as at least one Si membrane or layer surface or interface produces states that move the Fermi level sufficiently to enable the thermal excitation of 'bulk' carriers in the Si. In the present case that is a surface clean enough to produce the dimer reconstruction and surface electronic bands, but many other possibilities suggest themselves. For example, organic thin films can be tailored to have bandgaps with a lowest unoccupied molecular orbital or conduction band close to the Si valence band. The Si membrane layer will become hole doped, exactly as we describe here for a clean surface. The conduction properties of the organic layer (or whatever interface layer is created) are irrelevant. In addition to organic-inorganic interfaces, epitaxial insulators or insulators that form a disordered interface with Si provide practical examples of situations in which bulk conduction may be enabled by states at a surface or interface.

Conduction will be possible in very thin layers or membranes of other materials when the surface Fermi level is pinned far from mid-bandgap by surface or interface states. Considerations of the density of bulk dopants become irrelevant as long as the doping is not degenerate. For materials that have inappropriate surface states, such as cleaved GaAs(110), measurements that depend on electron transport may fail in thin enough films or membranes if interface trap states deplete the bulk carriers. If a material is deposited that provides these states, the conduction becomes enabled.

As we have described, in our clean-Si layers, conduction through the Si layer is enabled by interaction with the surface states created by reconstruction, as demonstrated with STM measurements. STM imaging will be possible on SOI(001) at any Si layer thickness at which the surface bands are well developed. In our situation, disruption of the surface bands will prevent STM imaging (or other forms of imaging requiring electron transport) of the thin SOI(001). Disruption of the π -bonded dimer chains that produce the surface bands can come, for example, via introduction of surface structural disorder, chemisorption of H (ref. 23) and other gases²⁴, or oxidation^{22,25}. Previous efforts to image SOI(001) with STM have encountered contamination and apparent disruption of these surface states, removing the surface state-bulk interaction that enables high-mobility conduction^{8,9}. The surface state-bulk interaction could be preserved outside vacuum using approaches that leave the surface chemically protected, but preserve the dimer structural motif and the high density of surface states^{26,27}.

In materials with appropriate surface or interface bands to allow thermal interaction with bulk bands, arbitrarily thin membranes or layers should show electronic conduction. Electronic-transport measurements (including, but not limited to, STM) should permit study of the transition from thin but still conventional layer structures to potentially novel phases created by the interaction of front and back layer interfaces.

METHODS

Surface preparation of ultrathin SOI for STM imaging consists of *ex situ* and *in situ* cleaning. *Ex situ*, the SOI is triple IMEC cleaned²⁸. A final 'Piranha' (mixture of H₂SO₄ and H₂O₂, 2–4:1) clean is performed to terminate the surface with a thin (1–2 nm) oxide for protection as the sample is transported to the chamber. The samples are introduced into an ultrahigh-vacuum STM, with a base pressure below 1×10^{-10} torr. Traditional *in situ* preparation of surfaces by direct heating to 1,500 K for several minutes used for bulk Si is not possible for SOI because the Si template layer dewets. Instead, we slowly deposit several ML of Si or Ge at 700 °C at 0.5 ML min^{−1} to remove the oxide. Finally we flash the sample to 800 °C (below the critical temperature for Si film dewetting) for two minutes, quench and anneal it at 600 °C for 30 min, radiation cool it, and transfer it to our STM chamber.

Received 21 March; accepted 28 November 2005.

1. Lu, Z. H., Lockwood, D. J. & Baribeau, J. M. Quantum confinement and light-emission in SiO₂/Si superlattices. *Nature* **378**, 258–260 (1995).
2. Legrand, B., Agache, V., Nys, J. P., Senez, V. & Stievenard, D. Formation of silicon islands on a silicon on insulator substrate upon thermal annealing. *Appl. Phys. Lett.* **76**, 3271–3273 (2000).
3. Liu, W. & Asheghi, M. Phonon-boundary scattering in ultrathin single-crystal silicon layers. *Appl. Phys. Lett.* **84**, 3819–3821 (2004).
4. Liu, F., Huang, M. H., Rugheimer, P. P., Savage, D. E. & Lagally, M. G. Nanostressors and the nanomechanical response of a thin silicon film on an insulator. *Phys. Rev. Lett.* **89**, 136101 (2002).
5. Celler, G. K. & Cristoloveanu, S. Frontiers of silicon-on-insulator. *J. Appl. Phys.* **93**, 4955–4978 (2003).
6. Kim, Y. A 0.25 mm 600MHz 1.5V SOI 64b ALPHATM microprocessor. *ISSCC Dig. Tech. Pap. I*, 432–433 (1999).
7. Yan, R. H., Ourmazd, A. & Lee, K. F. Scaling the Si MOSFET — From bulk to SOI to bulk. *IEEE Trans. Electron. Dev.* **39**, 1704–1710 (1992).
8. Sutter, P., Ernst, W. & Sutter, E. Scanning tunneling microscopy on ultrathin silicon on insulator (100). *Appl. Phys. Lett.* **85**, 3148–3150 (2004).
9. Lin, K. C., Holland, O. W., Feldman, L. C. & Weitering, H. H. Surface characterization of silicon on insulator material. *Appl. Phys. Lett.* **72**, 2313–2315 (1998).
10. Hovel, H. et al. Status of 300 mm SOI material: comparisons with 200 mm. *Proc. Electrochem. Soc.* **2003–05**, 75–80 (2003).
11. Swartzentruber, B. S., Kitamura, N., Lagally, M. G. & Webb, M. B. Behavior of steps on Si(001) as a function of vicinality. *Phys. Rev. B* **47**, 13432–13441 (1993).
12. Tromp, R. M., Hamers, R. J. & Demuth, J. E. Si(001) dimer structure observed with scanning tunneling microscopy. *Phys. Rev. Lett.* **55**, 1303–1306 (1985).
13. Fontes, E., Patel, J. R. & Comin, F. Direct measurement of the asymmetric dimer buckling of Ge on Si(001). *Phys. Rev. Lett.* **70**, 2790–2793 (1993).
14. Schroder, D. K. *Semiconductor Material and Device Characterization* (Wiley & Sons, New York, 1998).
15. McWhorter, P. J. & Winokur, P. S. Simple technique for separating the effects of interface traps and trapped-oxide charge in metal-oxide-semiconductor transistors. *Appl. Phys. Lett.* **48**, 133–135 (1986).
16. Cristoloveanu, S. & Sheng, L. *Electrical Characterization of Silicon-on-Insulator Materials and Devices* (Kluwer, Boston, 1995).
17. Northrup, J. E. Electronic structure of Si(100)C(4x2) calculated within the GW approximation. *Phys. Rev. B* **47**, 10032–10035 (1993).
18. Strosio, J. A. & Ho, W. Long-range quasielastic scattering of low-energy electrons by conduction-band surface-plasmons on Si(111)7x7. *Phys. Rev. Lett.* **54**, 1573–1576 (1985).
19. Heike, S., Watanabe, S., Wada, Y. & Hashizume, T. Electron conduction through surface states of the Si(111)-(7x7) surface. *Phys. Rev. Lett.* **81**, 890–893 (1998).
20. Tanikawa, T., Matsuda, I., Kanagawa, T. & Hasegawa, S. Surface-state electrical conductivity at a metal-insulator transition on silicon. *Phys. Rev. Lett.* **93**, 016801 (2004).
21. Petersen, C. L., Grey, F., Shiraki, I. & Hasegawa, S. Micro-four-point probe for studying electronic transport through surface states. *Appl. Phys. Lett.* **77**, 3782–3784 (2000).
22. Yoo, K. & Weitering, H. H. Electrical conductance of reconstructed silicon surfaces. *Phys. Rev. B* **65**, 115424 (2002).
23. Chen, D. X. & Boland, J. J. Chemisorption-induced disruption of surface electronic structure: Hydrogen adsorption on the Si(100)-2x1 surface. *Phys. Rev. B* **65**, 165336 (2002).
24. Chen, D. X. & Boland, J. J. Spontaneous roughening of low-coverage Si(100)-2x1: Cl surfaces: Patch formation on submonolayer halogenated surface. *Phys. Rev. B* **70**, 205432 (2004).
25. Hollinger, G. & Himpsel, F. J. Oxygen chemisorption and oxide formation on Si(111) and Si(100) surfaces. *J. Vac. Sci. Technol. A* **1**, 640–645 (1983).
26. Hamers, R. J. et al. Cycloaddition chemistry of organic molecules with semiconductor surfaces. *Acc. Chem. Res.* **33**, 617–624 (2000).
27. Seino, K., Schmidt, W. G. & Bechstedt, F. Organic modification of surface electronic properties: A first-principles study of uracil on Si(001). *Phys. Rev. B* **69**, 245309 (2004).
28. Meuris, M. et al. The IMEC clean — A new concept for particle and metal removal on Si surfaces. *Solid State Technol.* **38**, 109–114 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This research was supported by the US National Science Foundation, the US Department of Energy, and the US Air Force Office of Scientific Research.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.G.L. (lagally@engr.wisc.edu).

Initial corrosion observed on the atomic scale

F. U. Renner^{1,2}, A. Stierle¹, H. Dosch¹, D. M. Kolb³, T.-L. Lee² & J. Zegenhagen²

Corrosion destroys more than three per cent of the world's GDP¹. Recently, the electrochemical decomposition of metal alloys has been more productively harnessed to produce porous materials with diverse technological potential^{2,3}. High-resolution insight into structure formation during electrocorrosion is a prerequisite for an atomistic understanding and control of such electrochemical surface processes. Here we report atomic-scale observations of the initial stages of corrosion of a Cu₃Au(111) single crystal alloy within a sulphuric acid solution. We monitor, by *in situ* X-ray diffraction with picometre-scale resolution, the structure and chemical composition of the electrolyte/alloy interface as the material decomposes. We reveal the microscopic structural changes associated with a general passivation phenomenon of which the origin has been hitherto unclear. We observe the formation of a gold-enriched single-crystal layer that is two to three monolayers thick, and has an unexpected inverted (CBA-) stacking sequence. At higher potentials, we find that this protective passivation layer dewets and pure gold islands are formed; such structures form the templates for the growth of nanoporous metals². Our experiments are carried out on a model single-crystal system. However, the insights should equally apply within a crystalline grain of an associated polycrystalline electrode fabricated from many other alloys exhibiting a large difference in the standard potential of their constituents⁴, such as stainless steel (see ref. 5 for example) or alloys used for marine applications, such as CuZn or CuAl.

A common macroscopic fingerprint of the various electrochemical processes that occur during corrosion is provided by so-called cyclic voltammograms (current–voltage curves). Such a cyclic voltammogram for a pristine Cu₃Au(111) surface is shown in Fig. 1a, which is typical for the corrosion behaviour of alloys consisting of components with different noble character^{6–8}. A general observation is that—for the initial cycle (performed on the pristine sample)—the current exhibits a pronounced maximum, here in the potential regime between 100 mV and 300 mV (Cu dissolution starts at about 100 mV), which is not observed in subsequent cycles of the voltammogram. This phenomenon has given rise to many speculations about the possible formation of an ultrathin passivation layer. Recent *in situ* electrochemical scanning tunnelling microscopy experiments in this potential regime of passivation revealed the formation of mono-atomic deep vacancies, which have been interpreted as initial dissolution of the more reactive element from Ag–Au (ref. 9) and Cu–Au (refs 10–12) alloy surfaces. Thickness and composition of the passivation layer have been inferred only indirectly from the voltammograms¹³.

Above a critical potential E_c the current increases sharply owing to a bulk de-alloying process that produces nanoporous layers of the nobler component. Such a de-alloying process has been observed for binary alloys like Cu–Pd (ref. 14), Ag–Au (refs 2, 15, 16) and Cu–Au (ref. 16). *Ex situ* transmission electron microscopy studies of Cu–Au, Cu–Zn and Ag–Au alloys show evidence for the presence of either pits or three-dimensional islands with a typical size of 5–10 nm

(refs 16–19). For controlled production of technically relevant ‘corrosion’ products like nanoporous materials or related etching patterns, an understanding of the relevant parameters for the formation of patterns with regular morphology below and above the critical potential E_c is necessary. A detailed knowledge of the

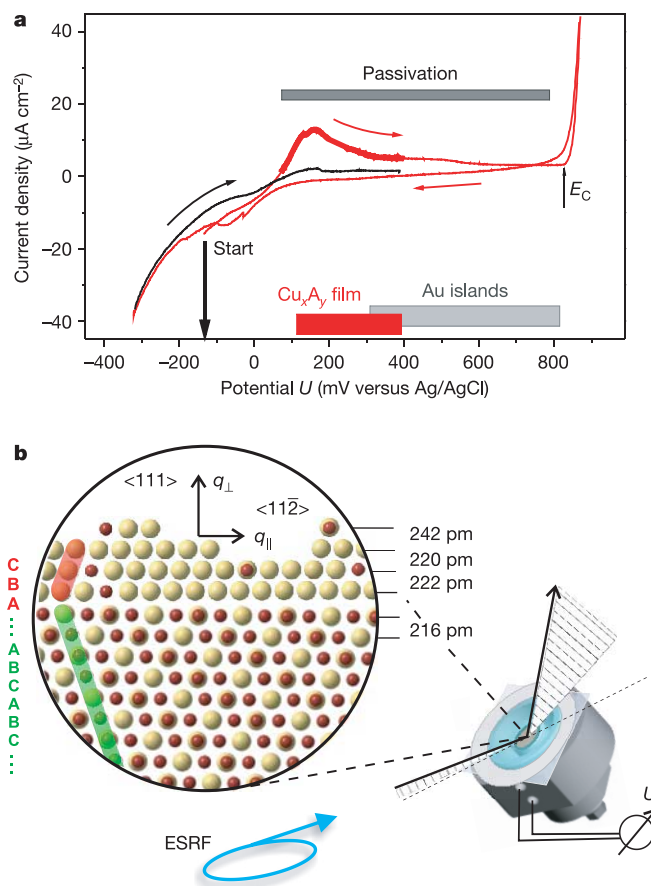


Figure 1 | Voltammogram and passivation layer formation.

a, Voltammograms recorded at 5 mV s^{-1} show the typical behaviour for corrosion passivation: the red line indicates the virgin current–voltage curve, exhibiting a pronounced maximum between 100 and 300 mV, which is not present in the subsequent cycle (black curve). The bold red curve marks the potential regime in which the X-ray measurements have been performed. Before the passivation breakdown at E_c a current plateau is observed. **b**, Structural model of the ultrathin passivation layer resulting from the fit to the X-ray diffraction data. Au atoms are represented by yellow spheres, and Cu atoms by red spheres. The ABC stacking of the substrate is inverted in the single crystalline overlayer. The two topmost layers are only partially occupied. A schematic view of the *in situ* X-ray diffraction cell is included.

¹Max-Planck-Institut für Metallforschung, Heisenbergstrasse 3, D-70569 Stuttgart, Germany. ²European Synchrotron Radiation Facility, Boîte Postale 220, F-38043 Grenoble, France. ³Abteilung Elektrochemie, Universität Ulm, Albert-Einstein-Allee 47, D-89081 Ulm, Germany.

structure, composition and defects of the passivation layer is thus needed to understand the formation of the nanopores.

The *in situ* X-ray diffraction experiments on the Cu_3Au (111) surface have been performed in the potential regime of the initial Cu dissolution shown as the bold red line in Fig. 1a. The use of a single crystal enables us to observe with high resolution the initial dealloying process as it occurs on a crystalline grain of the associated polycrystalline electrode. By recording the intensity along the momentum transfer $q_{\parallel}$ in the surface plane, information on the in-plane orientation and lattice parameters of epitaxial overlayers is obtained. By varying $q_{\perp}$ (normal to the surface, 'rod scans') the full three-dimensional structure of the overlayer is derived²⁰ (displayed in Fig. 1b). We note that the (111) planes exhibit the average Cu_3Au composition, and thus the corrosion process described below should be independent of ordering and disordering phenomena in the alloy.

First, we exposed the pristine surface to the de-aerated 0.1 M H_2SO_4 electrolyte with the sample potential kept at -100 mV below the equilibrium potential for Cu dissolution²¹. After this initial exposure to the electrolyte, the in-plane scan along the $(1, -1, 0)$ direction exhibits only the $(2, -2, 0)$ bulk diffraction peak from the Cu_3Au substrate. When increasing the potential up to 300 mV, a new diffraction peak can be observed at $q_{\parallel}$ values smaller than for the

Cu_3Au $(2, -2, 0)$ diffraction peak (Fig. 2a). This new peak arises from a well-ordered, nanosized epitaxial film with an in-plane nearest-neighbour distance of $a = 280.6$ pm, which lies between that of the Cu_3Au substrate ($a = 265.2$ pm) and that of pure Au ($a = 288.4$ pm). Upon further increase of the potential, the growing peak intensity directly shows the lateral growth of the passivation layer. The observed peak shift towards the Au $(2, -2, 0)$ reflection reveals a simultaneous increase of the in-plane nearest-neighbour distance (the same behaviour was observed in all symmetry-equivalent directions). From the peak width a lateral domain size of about 20 nm is estimated. The in-plane orientational disorder of the film is typically $\pm 1^\circ$.

The three-dimensional structure of the passivation layer is obtained from a reciprocal space map in the $(1, 1, -2)/(1, 1, 1)$ plane (Fig. 2b). The broad diffraction features that arise between the Cu_3Au substrate peaks give direct information on the close-packed layer stacking and orientation of the passivation layer. From the peak width a thickness of three monolayers can be estimated. The $q_{\perp}$ positions of the overlayer intensity distribution with respect to the substrate Bragg reflections disclose an unexpected structural feature, a CBA stacking, which inverts the ABC stacking sequence of the Cu_3Au face-centred-cubic (f.c.c.)-substrate (we note that the peak positions of the overlayer are shifted in $q_{\perp}$ by $(1/3) \cdot (1, 1, 1)$). A possible explanation for this unexpected stacking inversion is the blocking of regular f.c.c. sites by adsorbates from the electrolyte.

At a potential of 270 mV, the crystallographic structure of the epitaxial layer has been determined by a quantitative measurement of the integrated intensities of several Bragg diffraction rods normal

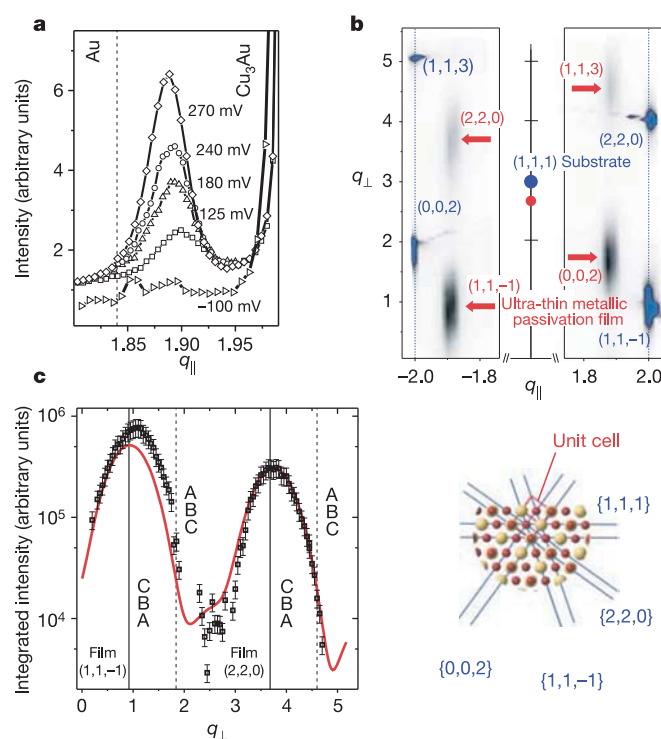


Figure 2 | In-situ structural characterization of the passivation layer.

a, Diffracted intensity along $q_{\parallel}$ in units of $(1, -1, 0)$ for varying electrochemical potential; the statistical error bar is below the symbol size. A Bragg peak from the epitaxial passivation layer is growing at $(1.89, -1.89, 0)$; in addition the $(2, -2, 0)$ substrate peak is observed. **b**, Experimental reciprocal space map in the $(1, 1, -2)/(1, 1, 1)$ plane characteristic for the potential regime from -100 mV to 270 mV. $q_{\parallel}$ is given in units of $1/3(1, 1, -2)$. The Bragg reflections from the passivation layer and the Cu_3Au substrate are observed and from the shift of the Bragg peak positions between the substrate and the film along $q_{\perp}$, it can be inferred that an epitaxial, single-crystalline overlayer with inverted CBA stacking is forming. **c**, Integrated intensities of the ultrathin passivation layer along $q_{\perp}$ ($q_{\parallel} = (1.89, -1.89, 0)$); the red line represents the best fit to the data based on the structural model depicted in Fig. 1b. Solid lines indicate the $q_{\perp}$ position for a layer with CBA stacking, dashed lines the $q_{\perp}$ position for an ABC stacking.

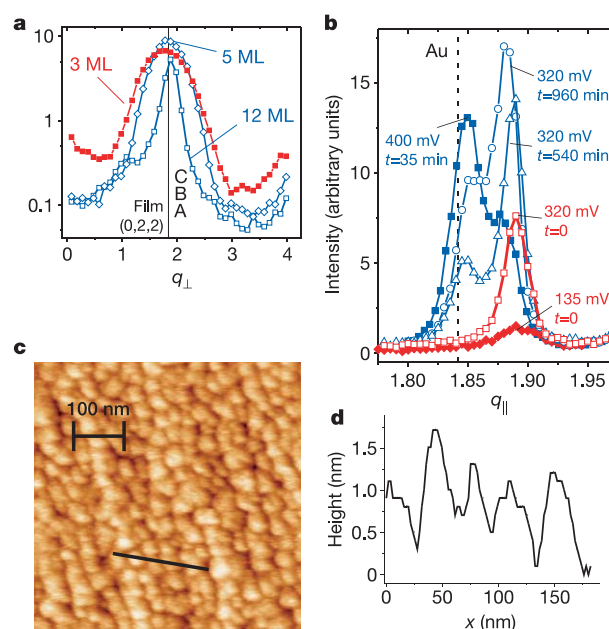


Figure 3 | Transformation of the passivation layer into Au islands. **a**, The ultrathin layer gives rise to a broad intensity distribution along $q_{\perp}$. With both in-plane peaks present, the pure gold peak in the corresponding $q_{\perp}$ -scan has a much narrower width, indicating an island height of typically 10–15 monolayers (2–3 nm). The ultrathin passivation layer is transforming into Au islands, maintaining predominantly the initial CBA stacking. **b**, By increasing the potential above 300 mV on the timescale of hours, an in-plane peak at a $q_{\parallel}$ position close to epitaxial gold is emerging in addition to the ultrathin passivation layer peak. **c**, *Ex situ* atomic force microscope image ($0.5 \mu\text{m} \times 0.5 \mu\text{m}$) after applying a potential of 450 mV versus Ag/AgCl; in this potential regime the formation of 2–3-nm-thick pure Au islands was observed in the X-ray diffraction experiments, which is a typical corrugation of the surface. **d**, Height along the line segment in **c**. The typical lateral dimension of the islands is 20 nm. The measured vertical corrugation of 1.5 nm is probably limited by tip effects.

to the surface. Figure 2c shows a rod together with the best fit (red curve), obtained from a simultaneous fit of three symmetry-inequivalent rods. The structural model derived from the fit is depicted in Fig. 1b: the film consists of three hexagonally close-packed Au-enriched planes, with an inverted f.c.c. CBA stacking. The film is single-crystalline and, furthermore, does not show lateral coexistence of ABC and CBA stacked domains, which would be expected for the hetero-epitaxial growth of f.c.c. materials in the (111) direction. This implies that the stacking information from the substrate is transmitted to the film. The structural model includes two fully occupied layers, a third layer with 45% occupancy and a fourth layer with 5% occupancy. This atomic 'roughness' gives rise to a strong damping of the so-called X-ray Laue oscillations which would otherwise arise around the Bragg intensities in the direction normal to the thin film. The interlayer spacing has increased by 2% compared to the Cu_3Au bulk value of 217 pm.

For potentials between 300 mV and the critical potential at 800 mV the corrosion behaviour changes dramatically. We find that the initial passivation layer transforms into 2.6-nm-thick (12 monolayers) Au-rich islands, which inherit the inverted CBA stacking sequence. This observation can directly be inferred from the diffraction maxima along $q_{\perp}$ associated with the CBA stacking (Fig. 3a): the formation of Au-rich islands, which takes place after slowly increasing the potential up to 400 mV over a period of 20 h, gives rise to additional Bragg intensity, which exhibits a narrower lineshape perpendicular to the surface and a Bragg signal shifted in-plane ($q_{\parallel}$) towards the Au (2, -2, 0) reflection (Fig. 3b). While the Bragg peak intensity of these islands grows with increasing potential, the peak intensity of the ultrathin passivation layer decreases and finally vanishes. This transition regime from the dense passivation layer to islands of the more noble component has recently been predicted by kinetic Monte Carlo simulations²².

Ex situ atomic force microscope images of the Cu_3Au (111) surface after applying a potential >300 mV always show a surface densely covered with islands of a height >1.5 nm and 20 nm lateral dimension (Fig. 3c and d). These islands can thus be directly associated with the Au islands observed in the X-ray diffraction experiment. The Au islands exhibit a weak, but significant, short-range correlation and a homogeneous distribution of island sizes and distances that could be reminiscent of a spinodal decomposition process during island formation.

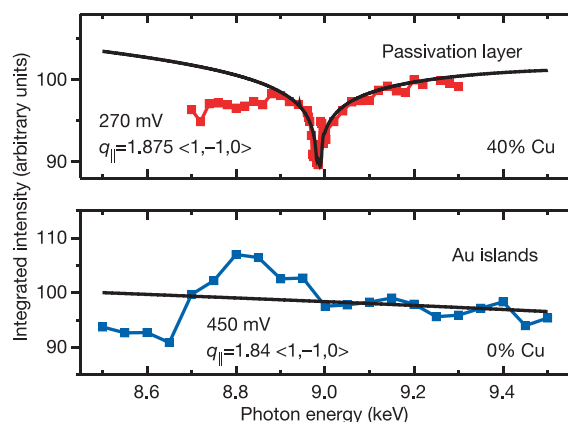


Figure 4 | Integrated anomalous in-plane diffraction intensities of the ultrathin passivation layer, and Au islands. The solid lines represent model calculations with varying Cu concentration. For the diffraction intensity of the passivation layer (top) a dip at the copper K edge is clearly visible owing to the contribution of the anomalous dispersion f'' and absorption f'' correction factors to the atomic form factor of Cu, and the model calculations agree best for a Cu concentration of 40%. The signal from the Au islands (bottom) exhibits no contribution of copper to the diffracted intensity.

To get access to the copper concentration in the ultrathin passivation layer and in the Au islands, energy-dependent Bragg diffraction has been performed in the energy range of the Cu absorption K edge (8.979 keV). In Fig. 4 this anomalous Bragg intensity associated with the ultrathin passivation layer and the Au islands is plotted together with model calculations. For the ultrathin passivation layer a dip is observed at the position of the Cu K edge, and from the model calculation a copper concentration of $40 \pm 10\%$ can be determined. The anomalous diffraction peaks from the thicker islands exhibit no such dip and thus appear to be almost pure Au, in agreement with the observed shift of the island in-plane lattice constant towards the bulk Au value. A small deviation from the exact lattice constant of pure Au indicates that some residual strain is present and that small amounts of Cu may still be present in the islands formed at potentials above 300 mV.

This combined structural and chemical *in situ* analysis of the growth of ultrathin passivation layers allows us to understand the initial electrochemical corrosion of alloys on a microscopic level. At potentials up to 300 mV, the initial Cu dissolution leads to an ultrathin, single-crystalline f.c.c. Au-Cu alloy layer, which acts as a nanosized protection layer against further dealloying. At potentials above 300 mV, almost pure Au islands start to grow, at the expense of the initial Au-Cu alloy layer, maintaining the original, inverted CBA stacking sequence of the Au-Cu layer.

METHODS

Sample preparation. The Cu_3Au single crystal was grown in the Czochralski apparatus of the crystal growth facility of the Max-Planck-Institut für Metallforschung (MPI-MF). After mechanical polishing, the sample surface was treated by several sputter-annealing cycles under ultrahigh vacuum conditions and the surface cleanliness and the L_{12} superstructure ordering was verified with Auger electron spectroscopy and low-energy electron diffraction. To obtain a well-ordered superstructure, the sample was kept at a temperature 30 K below the order-disorder phase transition temperature of Cu_3Au for typically 12 h. The high quality of the surface superstructure was confirmed by surface X-ray diffraction²³. Because the corrosion of the sample surface is an irreversible process, this preparation was repeated for each experiment. After ultrahigh-vacuum preparation and characterization, the sample was transferred in air into the electrochemical cell for the *in situ* X-ray investigation.

Electrochemical cell for *in situ* X-ray diffraction. Besides the Cu_3Au working electrode, the cell is equipped with a Pt counter electrode and a Ag/AgCl miniature reference electrode. It is sealed with a 6- μm -thick Mylar foil, allowing entrance and exit of the X-ray beam^{24,25}. For the measurements of diffracted intensities the Mylar foil was slightly pressed on the sample surface by applying an under-inflation to the electrolyte in the cell. In this way the absorption of X-ray intensity can be minimized, despite a grazing incidence angle of less than 1° , while leaving a micrometre-thick layer of electrolyte above the surface and thus keeping the surface immersed and under potential control. Before changing the electrode potential, the electrolyte pressure was increased, providing a several-millimetre-thick solution layer above the sample surface.

X-ray diffraction. Cu_3Au is an ordered alloy in the cubic f.c.c.-like L_{12} structure ($a_0 = 375$ pm). All measurements are expressed in bulk reciprocal lattice coordinates of the Cu_3Au substrate. The surface X-ray diffraction measurements were carried out using z-axis and 2 + 2 diffractometers equipped with point detectors at beamline ID32 at the European Synchrotron Radiation Facility (ESRF) and the beamline BW2 at the Hamburger Synchrotronstrahlungslabor (HASYLAB at DESY). The measurements were performed with an incident angle close to the critical angle for total external reflection of Cu_3Au . Integrated intensities were obtained from rocking scans around the sample surface normal, after background subtraction and application of standard correction factors. The fitting of the integrated intensities was performed using the software package ROD²⁶. In anomalous diffraction experiments we make use of the fact that the dispersion correction f' to the atomic form factor f gets negative close to an X-ray absorption edge, which can significantly reduce the real part of f : $\text{Re}(f) = f_0 + f'$. The measurements were corrected for the energy dependence of the optical transmission function of the sample and for the energy-dependent absorption of the Mylar foil and the electrolyte.

Received 30 April; accepted 10 November 2005.

- Koch, G. H., Brongers, M. P. H., Thompson, N. G., Virmani, Y. P. & Payer, J. H. *Corrosion Cost and Preventive Strategies in the United States*. Report

- FHWA-RD-01-156 (Report by CC Technologies Laboratories, Inc. to Federal Highway Administration (FHWA), Office of Infrastructure Research and Development, McLean, 2001); (<http://www.corrosioncost.com/pdf/main.pdf>).
2. Erlebacher, J., Aziz, M. J., Karma, A., Dimitrov, M. & Sieradzki, K. Evolution of nanoporosity in dealloying. *Nature* **410**, 450–453 (2001).
 3. Weissmüller, J. *et al.* Charge-induced reversible strain in a metal. *Science* **300**, 312–315 (2003).
 4. Al-Kharafi, F. M., Ateya, B. G. & Allah, R. M. Selective solution of brass in salt water. *Appl. Electrochem.* **34**, 47–53 (2004).
 5. Williams, D. E., Newman, R. C., Song, Q. & Kelly, R. G. Passivity breakdown and pitting corrosion of binary alloys. *Nature* **350**, 216–219 (1991).
 6. Gerischer, H. in *Korrosion* Vol. 14 (ed. Heumann, T.) 59–64 (Chemie, Weinheim, 1962).
 7. Pickering, H. W. Characteristic features of alloy polarization curves. *Corr. Sci.* **23**, 1107–1120 (1983).
 8. Kaiser, H. & Eckstein, G. A. in *Corrosion of Alloys—Encyclopedia of Electrochemistry* (eds Stratmann, M. & Frankel, G. S.) Vol. 4, 156–186 (Wiley, Weinheim, 2003).
 9. Oppenheim, I. C., Trevor, D. J., Chidsey, Ch. E. D., Trevor, P. L. & Sieradzki, K. *In-situ* scanning tunnelling microscopy of corrosion of silver-gold alloys. *Science* **254**, 687–689 (1991).
 10. Chen, S. J. *et al.* Selective dissolution of copper from Au-rich Cu-Au alloys: an electrochemical study. *Surf. Sci.* **292**, 289–297 (1993).
 11. Eckstein, G. A. *In-situ Rastertunnelmikroskopische Untersuchungen zur selektiven Korrosion von niedrigindizierten Au₃Cu(hkl) und Cu₃Au(hkl) Legierungseinkristallen*. Dissertation, Erlangen (2001).
 12. Stratmann, M. & Rohwerder, M. A pore view of corrosion. *Nature* **410**, 420–423 (2001).
 13. Moffat, T. P., Fan, F.-R. F. & Bard, A. J. Electrochemical and scanning tunneling microscopic study of dealloying of Cu₃Au. *J. Electrochem. Soc.* **138**, 3224–3235 (1991).
 14. Kabius, B., Kaiser, H. & Kaesche, H. in *Surfaces Inhibition and Passivation* (eds McCafferty, E. & Brodd, R. J.) 562–573 (The Electrochemical Society, Pennington, 1986).
 15. Li, R. & Sieradzki, K. Ductile-brittle transition in random porous Au. *Phys. Rev. Lett.* **68**, 1168–1171 (1992).
 16. Pickering, H. W. & Swann, P. R. Electron metallography of chemical attack upon some alloys susceptible to stress corrosion cracking. *Corrosion* **19**, 373t–389t (1963).
 17. Forty, A. J. Corrosion micromorphology of noble metal alloys and depletion gilding. *Nature* **282**, 597–598 (1979).
 18. Forty, A. J. & Rowlands, G. A. Possible model for corrosion pitting and tunneling in noble metal alloys. *Phil. Mag. A* **43**, 171–188 (1980).
 19. Pickering, H. W. Formation of new phases during anodic dissolution of Zn-rich Cu-Zn alloys. *J. Electrochem. Soc.* **117**, 8–15 (1970).
 20. Robinson, I. K. & Tweet, D. J. Surface X-ray diffraction. *Rep. Prog. Phys.* **55**, 599–651 (1992).
 21. Zoubov, N., Vanleugenhage, C. & Pourbaix, M. in *Atlas d'Equilibres Electrochimiques* (ed. Pourbaix, M.) Ch. 14 (Gauthier Villars, Paris, 1963).
 22. Erlebacher, J. An atomistic description of dealloying. *J. Electrochem. Soc.* **151**, C614–C626 (2004).
 23. Stierle, A. *et al.* Dedicated Max-Planck beamline for the in-situ investigation of interfaces and thin films. *Rev. Sci. Instrum.* **75**, 5302–5307 (2004).
 24. Fleischmann, M. & Mao, B. W. In-situ X-ray diffraction studies of Pt electrode/solution interfaces. *J. Electroanal. Chem.* **229**, 125–139 (1987).
 25. Zegenhagen, J. *et al.* X-ray diffraction study of a semiconductor/electrolyte interface: n-GaAs(001)/H₂SO₄/(Cu). *Surf. Sci.* **352–354**, 346–351 (1996).
 26. Vlieg, E. ROD: a program for surface X-ray crystallography. *J. Appl. Crystallogr.* **33**, 401–405 (2000).

Acknowledgements We thank B. Krause, A. Reicho, S. Warren, B. C. C. Cowie, S. Thiess, O. Bunk and W. Drube for their help with the synchrotron measurements, and F. D'Anca and R. Felici for collaboration in developing the *in-situ* cell and installing the electrochemistry laboratory at INFM-OGG/ESRF. The Bundesministerium für Forschung und Bildung (BMBF) is acknowledged for financial support.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to F.U.R. (renner@esrf.fr).

Winter forest soil respiration controlled by climate and microbial community composition

Russell K. Monson^{1,2*}, David L. Lipson^{4*}, Sean P. Burns^{1,5}, Andrew A. Turnipseed⁵, Anthony C. Delany⁵, Mark W. Williams³ & Steven K. Schmidt¹

Most terrestrial carbon sequestration at mid-latitudes in the Northern Hemisphere occurs in seasonal, montane forest ecosystems¹. Winter respiratory carbon dioxide losses from these ecosystems are high, and over half of the carbon assimilated by photosynthesis in the summer can be lost the following winter^{2,3}. The amount of winter carbon dioxide loss is potentially susceptible to changes in the depth of the snowpack; a shallower snowpack has less insulation potential, causing colder soil temperatures and potentially lower soil respiration rates. Recent climate analyses have shown widespread declines in the winter snowpack of mountain ecosystems in the western USA and Europe that are coupled to positive temperature anomalies^{4–6}. Here we study the effect of changes in snow cover on soil carbon cycling within the context of natural climate variation. We use a six-year record of net ecosystem carbon dioxide exchange in a subalpine forest to show that years with a reduced winter snowpack are accompanied by significantly lower rates of soil respiration. Furthermore, we show that the cause of the high sensitivity of soil respiration rate to changes in snow depth is a unique soil microbial community that exhibits exponential growth and high rates of substrate utilization at the cold temperatures that exist beneath the snow. Our observations suggest that a warmer climate may change soil carbon sequestration rates in forest ecosystems owing to changes in the depth of the insulating snow cover.

The recent global proliferation of tower-measurement networks has made it possible to analyse details of the coupling between climate dynamics and the carbon (C) cycle⁷. Most past studies have focused on ecosystem–atmosphere CO₂ exchange during the growing season because the instantaneous flux rates are so much higher than during colder periods. In seasonal forests, however, small but

continuous rates of ecosystem respiration during the winter can, in some cases, completely determine annual rates of C sequestration^{2,3}. Studies using artificial snow removal have demonstrated a critical role for thermal insulation in determining winter biogeochemical cycling^{8–10}. In this study, we aimed to move beyond experimental manipulations of snow cover, which can cause artificial treatment effects, and examine the factors that control winter soil C cycling within the context of natural climate variation.

We used the eddy covariance approach between 1 November 1998 and 31 October 2004 to continuously measure net ecosystem CO₂ exchange (NEE) at the Niwot Ridge Ameriflux site in the Rocky Mountains. Interannual variation in late-winter (1 March–15 April) cumulative NEE was not correlated with variation in mean air temperature ($P < 0.05$, Fig. 1a), but was correlated with variation in mean soil temperature (Fig. 1b) which, in turn, was correlated with variation in the 1 April snow-water equivalent (SWE), a measure of the cumulative winter snow pack (Fig. 1c). The relationship between mean soil temperature and SWE is best explained by a second-order polynomial, reflecting an asymptote at high snow depths as soil temperature approached its natural limit of 0 °C. The mean 1 April SWE for the period 1982–2004 at the Niwot Ridge site is 36.7 ± 3.3 cm, which falls just above the threshold of 31.4 cm, below which decreases in SWE affect late-winter NEE; nine of the past 23 years (39%) have been characterized by 1 April SWE values that fall below this threshold.

We developed a first-order exponential coefficient (R_T) to describe the temperature sensitivity of respiration (analogous to the Q_{10} coefficient used in biochemical studies). The interannual temperature dependency described by R_T was 6.6 when determined across all six years (Fig. 1b). The interannual range in late-winter NEE

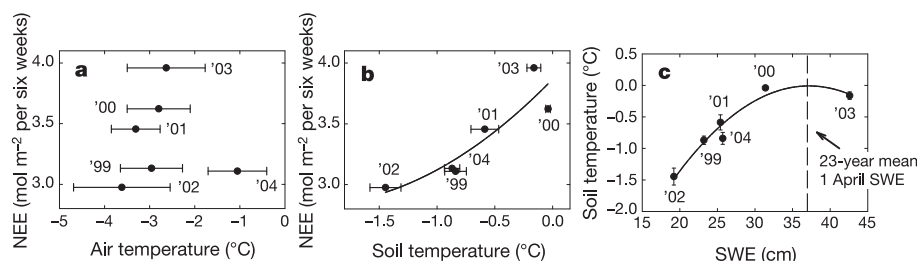


Figure 1 | Responses of NEE to air and soil temperature as influenced by SWE for the indicated years. a, The response of NEE to average daily air temperature. **b**, The response of NEE to average daily soil temperature ($y = a \exp^{bx}$, $a = 3.8244$, $b = 0.1963$, $R^2 = 0.81$, $P = 0.0152$). **c**, The

relationship between SWE (on 1 April) and average daily soil temperature ($y = ax^2 + bx + c$, $a = -0.0046$, $b = 0.339$, $c = -6.320$, $R^2 = 0.944$, $P = 0.0129$). Error bars are mean $\pm$ s.e. Data are for 1 March to 15 April.

¹Department of Ecology and Evolutionary Biology, ²Center for Interdisciplinary Research in Environmental Science, ³Department of Geography and Institute of Arctic and Alpine Research, University of Colorado, Boulder, Colorado 80309, USA. ⁴Department of Biology, San Diego State University, San Diego, California 92182, USA. ⁵National Center for Atmospheric Research, PO Box 3000, Boulder, Colorado 80305, USA.

*These authors contributed equally to this work.

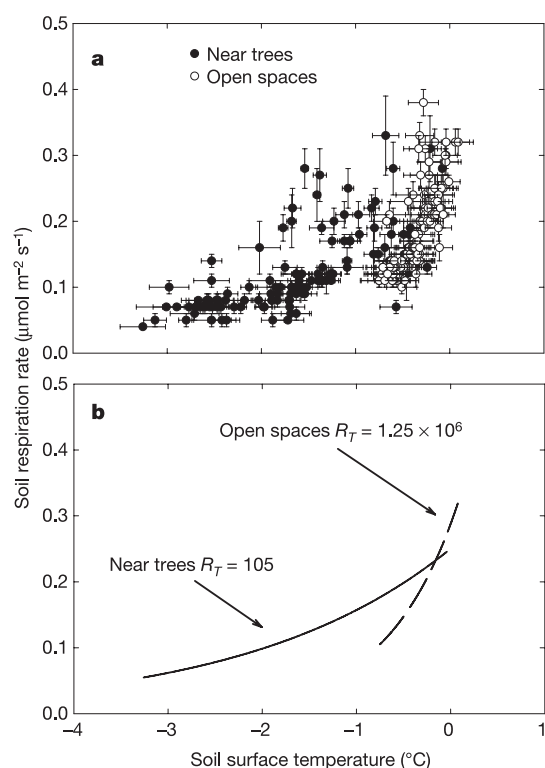


Figure 2 | The temperature sensitivity of soil respiration rate beneath the snow pack. **a**, The response of beneath-snow soil-respiration rate to soil temperature measured during the snow-covered period between 20 November 2003 and 22 March 2004 in open spaces between trees or in the spaces next to tree boles. Vertical and horizontal error bars represent the mean $\pm$ s.e. of four chambers located in open spaces and three chambers located next to trees. **b**, The mean temperature response and estimates of R_T from a first-order exponential model applied to the data in **a**.

evaluated across all six years ($0.99 \text{ mol C m}^{-2}$) represents 21% of the mean cumulative annual NEE for this site ($-4.81 \pm 0.8 \text{ mol C m}^{-2}$). Thus, the natural variance in late-winter snow pack and its effect on soil temperature has the potential to cause large variance in the capacity for annual C sequestration.

One past study has shown that high SWE coupled with persistent snow pack late into the spring stimulates photosynthetic gross

ecosystem CO₂ exchange (GEE), resulting in higher ecosystem CO₂ uptake². The results depicted in Fig. 1 demonstrate that when SWE is high late into the winter, before GEE has been stimulated, ecosystem respiration (R_e) is stimulated, resulting in higher CO₂ loss. The relative importance of these two opposing effects, as well as environmental conditions during the middle part of the growing season, will ultimately determine the NEE in any single year. In the six-year observation period of the present study, it was clear that springtime dynamics in GEE caused the largest interannual swings in cumulative annual NEE (ref. 2); however, late-winter differences in R_e were of sufficient magnitude to completely rearrange the ranking of years with respect to annual NEE, demonstrating the importance of spring SWE on annual NEE.

We made nearly continuous observations of soil respiration rate and soil temperature using an automated system. The measured CO₂ concentration was used to calculate the through-snow soil respiration rate. When data were pooled for the entire season, the calculated R_T was 105 and 1.25×10^6 , depending on whether measurements were made next to the boles of trees or in the open spaces between trees, respectively; these values are several orders of magnitude higher than the range of Q_{10} values (2–6) typically recorded for snow-free forest soils (Fig. 2)^{11–13}. When the analysis was restricted to observations between 0 and -1°C to achieve a temperature-normalized set of observations, the R_T was calculated as 104 for respiration near the boles of trees and 6.62×10^5 for respiration in the open spaces between trees.

Past studies of soils under laboratory conditions have also reported high temperature sensitivities for respiration at temperatures below 0°C ; though still several orders of magnitude below the highest values reported here¹⁴. One hypothesis that was proposed for the high temperature sensitivity invokes a physical limitation to substrate diffusion as liquid water disappears below 0°C . The disappearance of water between 0 and -1°C exhibits exponential behaviour, similar to what we observed for the temperature dependence of respiration¹⁵, and substrate limitations to microbial growth and respiration have been demonstrated in frozen soils¹⁶.

Our observation of lower R_T values for the soil respiration rate near tree boles, compared to open spaces, might be explained by differences in substrate diffusion limitations; substrate concentrations may be higher near the trees due to higher rates of rhizodeposition and litter inputs, and the higher concentration gradient for these substrates might overcome slow diffusion rates at subzero temperatures and render the overall process less sensitive to temperature. Substrate diffusion limitations would impose a

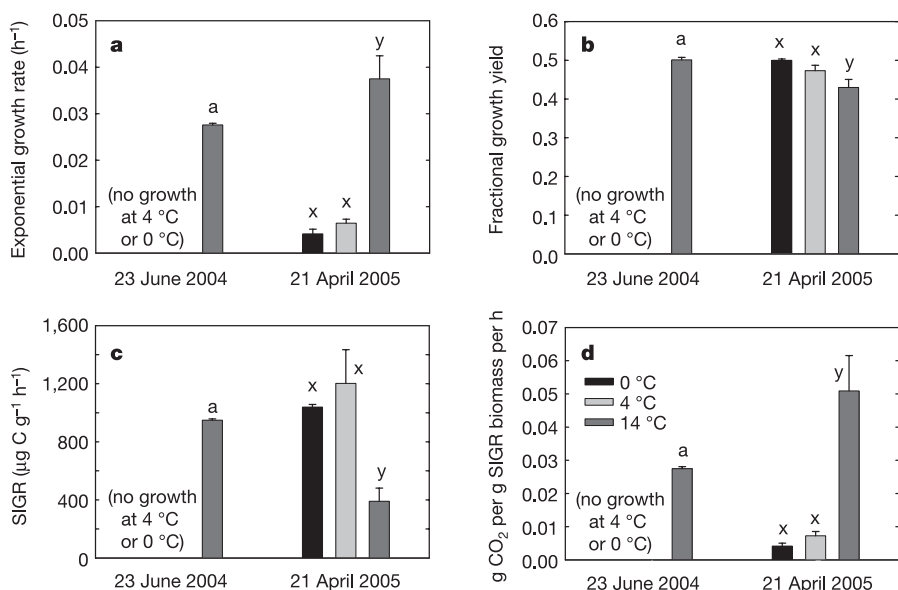


Figure 3 | Microbial growth kinetics in soils collected in summer and late winter, determined by SIGR experiments at temperatures ranging from 0 to 14°C . **a**, Exponential growth rate; **b**, growth yield; **c**, microbial biomass C capable of growth at the indicated temperature; **d**, rate of CO₂ production per unit of actively growing microbial biomass C. Different letters next to the bars of the same sample date indicate significant differences at $P < 0.05$ using Fisher's Protected Least Significant Difference Test. Vertical error bars represent s.e.

second-order constraint on respiratory metabolism and render first-order models inadequate. Whatever the limiting mechanism, the first-order exponential model driven by an R_T coefficient fits our observations well ($y = a \exp^{bx}$, $a = 0.287$, $b = 1.341$, $R^2 = 0.60$, $P < 0.0001$ for open spaces; $y = a \exp^{bx}$, $a = 0.250$, $b = 0.466$, $R^2 = 0.50$, $P < 0.0001$ for near trees). Thus, our conclusion that there is high R_T in the beneath-snow soil is supported by empirical regression.

We collected soils from beneath the spring snowpack and during the summer to determine whether the microbial communities differed in their seasonal composition, whether the microbial biomass was capable of increasing beneath the snow pack and whether microbial respiration responded to changes in substrate availability (Fig. 3). The summer microbial community was not capable of growth below 4°C. However, microbes collected from under the snowpack could grow exponentially at 0°C and their growth rates increased rapidly with increasing temperature. The growth yield (yield of biomass per unit of substrate assimilated) was highest at the lowest temperatures. These two parameters determine the rate of respiration per unit biomass and its response to temperature in a growing microbial population. The application of potassium glutamate substrate to soils collected from beneath the snowpack induced the largest substrate-induced growth respiration (SIGR) response at the lowest temperatures (Fig. 3c), and the rate of respiration per unit of SIGR biomass was especially sensitive to increases in temperature (Fig. 3d).

We constructed 16S ribosomal RNA clone libraries from the winter and summer soils after incubation in the glutamate SIGR experiments to identify which bacteria were responsible for the differential growth kinetics. The dominant bacteria in the winter library had sequences closely related to *Janthinobacterium* (38% of sequences) and, in summer, the sequences were closely related to *Burkholderia* (17% of the sequences). The overall DNA libraries were significantly different between the winter and summer soils by phylogenetic tail permutation (PTP) analysis ($P = 0.006$). These results show clearly that in addition to exhibiting different growth kinetic responses at low temperature, the soil microbial communities beneath the snowpack and during the summer are distinct in their taxonomic composition.

Long-term monitoring of mountain snow packs in the western USA and Europe have shown trends towards decreasing depth, with several mountain ranges experiencing 50–75% decreases, and these have been attributed to positive temperature anomalies^{4–6}. In other north-latitude ecosystems, climate interactions with local features of the landscape, including large lakes, are predicted to result in an increase in winter snow in the face of climate warming¹⁷. Trends in snow pack provide a critical, indirect coupling between climate and forest C cycles. Decreases in the winter snow pack will generally cause decreases in the loss of respired CO₂ from the soils of forest ecosystems, thus, enhancing the potential for soil C sequestration. The existence of a low-temperature adapted microbial flora living beneath the snow pack explains the connection between soil respiration rate and snow depth. These organisms function in a relatively narrow range of sub-zero soil temperatures, and are highly sensitive to warming or cooling outside this range. Our observations reveal that a warmer climate may change the beneath-snow soil temperature in forest ecosystems because of changes in the depth of the insulating snow cover, changing soil respiration rates and, concomitantly, soil C sequestration rates.

METHODS

Eddy covariance and beneath-snow respiration measurements. The Niwot Ridge AmeriFlux site is located at 3,050 m above sea level at 40°1'58" N; 105°32'47" W. Details of NEE and climate measurements can be found at http://urquell.colorado.edu/data_ameriflux/ (refs 13, 18, 19). The forest is dominated by lodgepole pine (*Pinus contorta*), subalpine fir (*Abies lasiocarpa*), and Engelmann spruce (*Picea engelmannii*). Soil temperatures were measured using copper-constantan thermocouples (1999–2001) or platinum resistance

thermometers (2001–2003) located within 50 m of the flux tower and in an area approximately 2 m from the nearest tree bole, and next to each of the soil chambers used for respiration measurements. Beneath-snow CO₂ concentrations were measured using a multi-inlet air-sampling system combined with a single LI-7000 IRGA gas analyser (LI-Cor Inc). Eleven chambers (2.5-litre volume) were placed on the ground before the first winter snow. An additional three inlets were placed above snow height to obtain the atmospheric CO₂ concentration. Air from the soil and air inlets was pumped (1.4 litres min⁻¹) to glass buffer volumes (2 litres) from which samples were drawn for analysis. Each buffer volume was sampled for 80 s every 30 min and only data from the final 20 s of the sampling time were used. Data on CO₂ concentration was continuously recorded at 1-s intervals during each sampling period. The IRGA was calibrated at 3–5-hour intervals. Calibration was obtained from three gases of a known CO₂ concentration (413, 504 and 749 p.p.m.v.) and calibrated accuracy to within ±2%. The maximum beneath-snow CO₂ concentration we observed was 3,295 p.p.m.v., which exceeded that for our highest calibration gas and introduced the possibility of error due to nonlinearity. We conducted analyses during the subsequent winter to determine the signal linearity across the range 0–3,000 p.p.m.v. using the original calibration gases plus a fourth mixture of 5,000 p.p.m.v. We calculated an upper-bounded, systematic error of ~5% at an observed value of 3,000 p.p.m.v. if the calibration was limited to the lower three gases only.

The flux of CO₂ through the snowpack was calculated using a steady-state diffusion model with driving variables being the difference in CO₂ concentration from soil to atmosphere, atmospheric pressure, air temperature, and snowpack depth, porosity and tortuosity^{20–24}. Snowpack physical characteristics were measured at 10-cm vertical intervals on a weekly basis after manually digging snow pits. SWE, snow temperature, grain type, size and snowpack stratigraphy were measured following previously published protocols and used to calculate snow density²⁵. Values of 1 April SWE were obtained from the US Natural Resources Conservation Service Snotel Database.

We explored the possibility that the CO₂ efflux was due to the release of ice-trapped air during soil warming. We assumed that the CO₂ efflux originates from the top 20 cm of the soil²⁶, that the maximum soil water content during the autumn is 30% (ref. 26), that the solubility of CO₂ in water is 0.35 g per 100 cm³, and we measured the pre-snow soil surface CO₂ partial pressure at 5.5×10^{-4} atm. We calculated that the CO₂ that could have been dissolved in autumnal soil water and released during the winter was 0.003 mol m⁻², which is less than 0.1% of the total observed CO₂ efflux (3.3 mol CO₂ m⁻²).

Values of R_T were calculated according to $R_T = (R_2/R_1)^{10/(T_2-T_1)}$, where R_2 and R_1 represent respiration rates measured at two different temperatures T_2 and T_1 , respectively. We note that R_T reflects responses to temperature, such as altered substrate supply and turnover of microbial populations, that are broader than the enzyme kinetics that drive traditional Q_{10} notation.

Microbial growth kinetics and substrate-use measurements. The growth kinetics of soil microbial communities were measured with the substrate-induced growth response method using potassium glutamate^{27,28}. Soils were incubated with sufficient substrate to induce growth, along with tracer levels of ¹⁴C-glutamate. Respired CO₂ was trapped in 1.0 M NaOH and counted by liquid scintillation. Exponential growth curves were fitted to respiration rate versus time²⁸. Experimental soils were frozen for later DNA extraction, polymerase chain reaction (PCR) amplification of the 16S rRNA gene, and clone library construction as described elsewhere²⁹. Sequences were either aligned manually or using the CLUSTAL algorithm within BioEdit. A maximum-likelihood tree was produced, and microbial communities were compared using the PTP test in PAUP. This method tests the null hypothesis that the samples are from the same community by comparing the original tree to a probability distribution made from 10,000 random permutations of the tree³⁰.

Received 1 July; accepted 20 December 2005.

- Schimmel, D., Kittel, T., Running, S., Monson, R., Turnipseed, A. & Anderson, D. Carbon sequestration studied in western US mountains. *Eos* **83**, 445–449 (2002).
- Monson, R. K. et al. Climatic influences on net ecosystem CO₂ exchange during the transition from wintertime carbon source to springtime carbon sink in a high-elevation, subalpine forest. *Oecologia* **146**, 130–147 (2005).
- Hubbard, R. M., Ryan, M. G., Elder, K. & Rhodes, C. C. Seasonal patterns in soil surface CO₂ flux under snow cover in 50 and 300 year old subalpine forests. *Biogeochemistry* **73**, 93–107 (2005).
- Mote, P. W., Hamlet, A. F., Clark, M. P. & Lettenmaier, D. T. Declining mountain snow pack in Western North America. *Bull. Am. Meteorol. Soc.* **86**, 39–49 (2005).
- Latenser, M. & Schneebeli, M. Long-term snow climate trends of the Swiss Alps (1931–99). *Int. J. Climatol.* **23**, 733–750 (2003).

6. Scherrer, S. C., Appenzeller, C. & Latenser, M. Trends in Swiss alpine snow days—The role of local and large-scale climate variability. *Geophys. Res. Lett.* **31**, L13215, doi:10.1029/2004GL020255 (2004).
7. Valentini, R. *et al.* Respiration as the main determinant of carbon balance in European forests. *Nature* **404**, 861–865 (2000).
8. Groffman, P. M. *et al.* Colder soils in a warmer world: A snow manipulation study in a northern hardwood forest ecosystem. *Biogeochemistry* **56**, 135–150 (2001).
9. Fitzhugh, R. D. *et al.* Soil freezing and the acid-base chemistry of soil solutions in a northern hardwood forest. *Soil Sci. Soc. Am.* **67**, 1897–1908 (2003).
10. Decker, K. L. M., Wang, D., Waite, C. & Scherbatskoy, T. Snow removal and ambient air temperature effects on forest soil temperatures in Northern Vermont. *Soil Sci. Soc. Am.* **67**, 1234–1242 (2003).
11. Lloyd, J. & Taylor, J. A. On the temperature dependence of soil respiration. *Funct. Ecol.* **8**, 315–323 (1994).
12. Kätterer, T., Reichstein, M., Andrén, O. & Lomander, A. Temperature dependence of organic matter decomposition: a critical review using literature data analyzed with different models. *Biol. Fertil. Soils* **27**, 258–262 (1998).
13. Monson, R. K. *et al.* Carbon sequestration in a high-elevation, subalpine forest. *Glob. Change Biol.* **8**, 1–20 (2002).
14. Mikan, C. J., Schimel, J. P. & Doyle, A. P. Temperature controls of microbial respiration in arctic tundra soils above and below freezing. *Soil Biol. Biochem.* **34**, 1785–1795 (2002).
15. Patterson, D. E. & Smith, M. W. The measurement of unfrozen water content by time domain reflectometry—results from laboratory tests. *Can. Geotech. J.* **18**, 131–144 (1981).
16. Brooks, P. D., McKnight, D. & Elder, K. Carbon limitation of soil respiration under winter snowpacks: potential feedbacks between growing season and winter carbon fluxes. *Glob. Change Biol.* **11**, 231–238 (2005).
17. Burnett, A. W., Kirby, M. E., Mullins, H. T. & Patterson, W. P. Increasing Great Lake-effect snowfall during the twentieth century: A regional response to global warming? *J. Clim.* **16**, 3535–3542 (2003).
18. Turnipseed, A. A., Anderson, D. E., Blanken, P. D., Baugh, W. M. & Monson, R. K. Airflows and turbulent flux measurements in mountainous terrain. Part 1. Canopy and local effects. *Agric. For. Meteorol.* **119**, 1–21 (2003).
19. Turnipseed, A. A., Anderson, D. E., Burns, S., Blanken, P. D. & Monson, R. K. Airflows and turbulent flux measurements in mountainous terrain. Part 2. Mesoscale effects. *Agric. For. Meteorol.* **125**, 187–205 (2004).
20. Sommerfeld, R. A., Mosier, A. R. & Musselman, R. C. CO₂, CH₄, and N₂O flux through a Wyoming snowpack. *Nature* **361**, 140–142 (1993).
21. Brooks, P. D., Schmidt, S. K. & Williams, M. W. Winter production of CO₂ and N₂O from alpine tundra: environmental controls and relationship to inter-system C and N fluxes. *Oecologia* **110**, 403–413 (1997).
22. Mast, M. A., Wickland, K. P., Striegl, R. T. & Clow, D. W. Winter fluxes of CO₂ and CH₄ from subalpine soils in Rocky Mountain National Park, Colorado. *Glob. Biogeochem. Cycles* **12**, 607–620 (1998).
23. Massman, W. J. *et al.* A model investigation of turbulence-driven pressure-pumping effects on the rate of diffusion of CO₂, N₂O, and CH₄ through layered snowpacks. *J. Geophys. Res.* **102**, 18851–18863 (1997).
24. Swanson, A. L., Lefer, B. L., Stroud, V. & Atlas, E. Trace gas emissions through a winter snowpack in the subalpine ecosystem at Niwot Ridge, Colorado. *Geophys. Res. Lett.* **32**, L03805, doi:10.1029/2004GL021809 (2005).
25. Williams, M. W., Brooks, P. D., Mosier, A. & Tonnessen, K. A. Mineral nitrogen transformations in and under seasonal snow in a high-elevation catchment, Rocky Mountains, USA. *Wat. Resour. Res.* **32**, 3175–3185 (1996).
26. Scott-Denton, L. E., Sparks, K. L. & Monson, R. K. Spatial and temporal controls over soil respiration rate in a high-elevation, subalpine forest. *Soil Biol. Biochem.* **35**, 525–534 (2003).
27. Lipson, D. A., Schmidt, S. K. & Monson, R. K. Links between microbial population dynamics and N availability in an alpine ecosystem. *Ecology* **80**, 1623–1631 (1999).
28. Schmidt, S. K. A substrate-induced growth-response (SIGR) method for estimating the biomass of microbial functional groups in soil and aquatic systems. *FEMS Microbiol. Ecol.* **101**, 197–206 (1992).
29. Lipson, D. A. & Schmidt, S. K. Seasonal changes in an alpine soil bacterial community in the Colorado Rocky Mountains. *Appl. Environ. Microbiol.* **70**, 2867–2879 (2004).
30. Martin, A. P. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Appl. Environ. Microbiol.* **68**, 3673–3682 (2002).

Acknowledgements We acknowledge the support provided by the US Department of Energy Terrestrial Carbon Program and National Institute for Global Environmental Change (NIGEC), the US National Science Foundation programmes on 'Long-Term Ecological Research', 'Ecology and Evolutionary Physiology', and 'Microbial Observatories', and the National Center for Atmospheric Research for support provided to the 'Carbon in the Mountains Experiment' (CME).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.K.M. (Russell.Monson@colorado.edu).

A basal tyrannosauroid dinosaur from the Late Jurassic of China

Xing Xu^{1,2}, James M. Clark³, Catherine A. Forster⁴, Mark A. Norell², Gregory M. Erickson⁵, David A. Eberth⁶, Chengkai Jia¹ & Qi Zhao¹

The tyrannosauroid fossil record is mainly restricted to Cretaceous sediments of Laurasia, although some very fragmentary Jurassic specimens have been referred to this group^{1,2}. Here we report a new basal tyrannosauroid, *Guanlong wucaii* gen. et sp. nov., from the lower Upper Jurassic of the Junggar Basin^{3,4}, northwestern China. *G. wucaii* is the oldest known tyrannosauroid and shows several unexpectedly primitive pelvic features^{5,6}. Nevertheless, the limbs of *G. wucaii* share several features with derived coelurosaurs^{7–9}, and it possesses features shared by other coelurosaurian clades¹⁰. This unusual combination of character states provides an insight into the poorly known early radiation of the Coelurosauria. Notably, the presumed predatory *Guanlong* has a large, fragile and highly pneumatic cranial crest that is among the most elaborate known in any non-avian dinosaur and could be comparable to some classical exaggerated ornamental traits among vertebrates.

Theropoda Marsh, 1881

Coelurosauria von Huene, 1914

Tyrannosauroida Osborn, 1905

Guanlong wucaii gen. et sp. nov.

Etymology. The generic name is derived from the Chinese *Guan* (crown) and *long* (dragon); the specific name is from the Chinese *wucui* (five colours), which refers to the rich colours of rocks that produced the specimens.

Holotype. IVPP (Institute of Vertebrate Paleontology and Paleoanthropology, Beijing) V14531 is a partly articulated skeleton preserving most of the elements.

Referred material. IVPP V14532 is much smaller than the holotype and is a fully articulated, nearly complete skeleton.

Locality and horizon. Wucaiwan area, Junggar Basin, Xinjiang; Oxfordian upper part of the Shishugou Formation^{3,4}.

Diagnosis. Medium-sized tyrannosauroid with the following autapomorphies: a deep and narrow groove along the anterior margin of the premaxilla; a distinct opening on the maxilla close to the premaxilla–maxilla contact; a complex, highly pneumatic nasal crest; a low, rugose ridge along the midline of the frontals; a dorsally flattened parietal with two parallel sagittal crests; a transverse ridge within the supratemporal fossa; a centropostzygapophyseal lamina on cervicodorsal vertebrae with its dorsal end expanding laterally; deep, longitudinal sulci on both ventral and dorsal surfaces of the distal caudal vertebrae (independently evolved in troodontids⁸); ventral part of scapular blade with sub-equilateral triangular cross-section and thick posterior margin; metacarpal II with prominent medioventral and laterodorsal processes proximally; manual phalanx II-2 with prominent medioventral process proximally; femoral greater trochanter much narrower anteroposteriorly than the lesser trochanter; distinct fossa on posterodistal surfaces of astragalus and

calcaneum; and pedal phalanx II-1 with prominent paired ventral processes proximally (Figs 1 and 2a, b; see also Supplementary Figs 1 and 2).

Both specimens (holotype and referred specimen) were collected from a tabular bed of finely laminated to massive, tan-coloured, tuffaceous mudstone, with characteristics indicating a paludal (wetlands) setting. Their preservational features indicate that IVPP V14532 died *in situ*, was possibly trampled at a later date by IVPP V14531, and was buried before subaerial disarticulation, and that IVPP V14531 was subaerially exposed for a significant time after death (see Supplementary Information for further analysis).

Description. The total body length of the *Guanlong* holotype is about 3 m. As in *Dilong*^{11,12}, a horizontal ridge is present along the ventral margin on the external surface of the jugal, the descending process of the squamosal is extremely long, and the posterior serrations are larger than the anterior ones on lateral teeth (also independently evolved in dromaeosaurids¹³). As in *Dilong*^{11,12} and alvarezsaurids, a row of foramina is located within a sharply defined groove on the external surface of the dentary. In troodontids, a similar groove is present, but is much wider and more ventrally positioned¹⁴. The most striking feature of *Guanlong* is a complex nasal crest consisting of a highly pneumatic median crest that is about 1.5 mm thick for most of its length, and four supporting lateral laminae. The mandible of *Guanlong* displays a relatively large external mandibular fenestra, an extremely weak surangular ridge and appears to lack a surangular foramen.

The cervical vertebrae are amphicoelous, with the anterior ones bearing axially long neural spines. The dorsal and sacral centra lack pneumatic openings.

The relative lengths of the forelimb and its components are similar to those of derived coelurosaurs^{8,9}, but are much longer than those of other tyrannosauroids^{1,15}. The length of the forelimb is only slightly less than 60% of the hindlimb length. The length of the humerus is more than 60% of the femoral length. The ulna and manus (which has a phalangeal formula of 2-3-4; Fig. 2d) are about 80% and 140% of the humeral length, respectively. Also similar to some derived coelurosaurs is the posteriorly bowed ulna and distally flattened radius with a slightly hooked and round distal margin^{7–9}. A tiny splint of bone attached to the proximal part of metacarpal III might represent a reduced metacarpal IV (Fig. 2d; see also Supplementary Fig. 2f), a feature also reported in *Ornitholestes*¹⁶ among the Coelurosauria. A distal carpal possesses a transverse trochlea proximally and a semilunate shape in ventral view (Fig. 2d; see also Supplementary Fig. 2f, g), representing the first known ‘semilunate’ carpal in an adult tyrannosauroid specimen¹. Interestingly, its position (mainly contacting metacarpal I) is similar to the condition in *Allosaurus*¹⁷ and oviraptorosaurs, whereas in therizinosauroids¹⁸, troodontids,

¹Institute of Vertebrate Paleontology and Paleoanthropology, Beijing 100044, China. ²American Museum of Natural History, New York, New York 10024, USA. ³Department of Biological Sciences, George Washington University, Washington DC 20052, USA. ⁴Department of Anatomical Sciences, Stony Brook University, Stony Brook, New York 11794, USA. ⁵Department of Biological Science, Florida State University, Tallahassee, Florida 32306, USA. ⁶Royal Tyrrell Museum, Drumheller, Alberta T0J 0Y0, Canada.

dromaeosaurids and basal birds⁹ it primarily articulates with metacarpal II. The latter condition is also consistent with some embryonic evidence from extant birds¹⁹. This suggests that a semilunate shape may have evolved on different carpals in different theropod groups, and thus caution should be used in identifying a semilunate carpal in phylogenetic analyses owing to uncertainties regarding homology.

The ilia are moderately inclined towards the midline but do not contact. The dorsal part of the pubis narrows to a thin sheet of bone posteriorly, which might be pierced by an obturator foramen (Fig. 2e). A relatively large foramen pierces the extremely thin sheet of bone on the anterodorsal part of the ischium, which is continuous with the thin sheet of bone down the shaft. Pedal digit I is attached to the posterior margin of metatarsal II (Fig. 2f), indicating a partially reversed hallux (see additional description in Supplementary Information).

A histological analysis was conducted to determine the longevity and developmental stages for the specimens. The results suggest that IVPP V14531 reached full adult size in 7 years and was a relatively old, late stationary stage individual at the time of death in its twelfth year of life (Fig. 2g). IVPP V14532 appears to have died during its sixth year of life. This animal was still actively growing and was in the exponential stage of development, as indicated by the same degree of

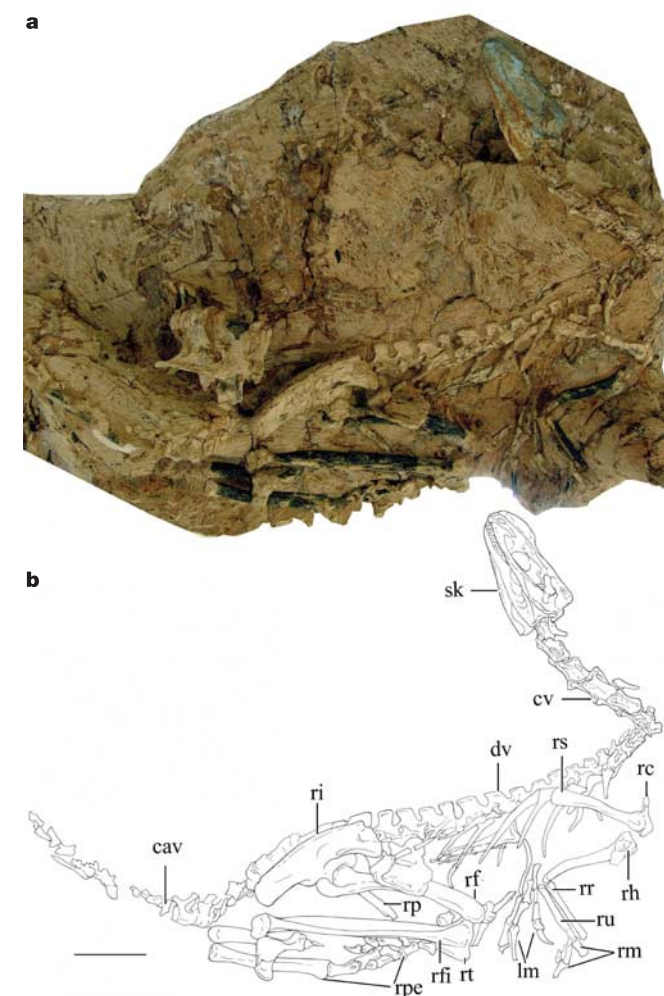


Figure 1 | *Guanlong wucaii* (IVPP V14532). **a**, Photograph, **b**, Line drawing of IVPP V14532. cav, caudal vertebrae; cv, cervical vertebrae; dv, dorsal vertebrae; lm, left manus; rc, right coracoid; rf, right femur; rfi, right fibula; rh, right humerus; ri, right ilium; rm, right manus; rp, right pubis; rpe, right pes; rr, right radius; rs, right scapula; rt, right tibia; ru, right ulna; sk, skull. Scale bar, 8 cm.

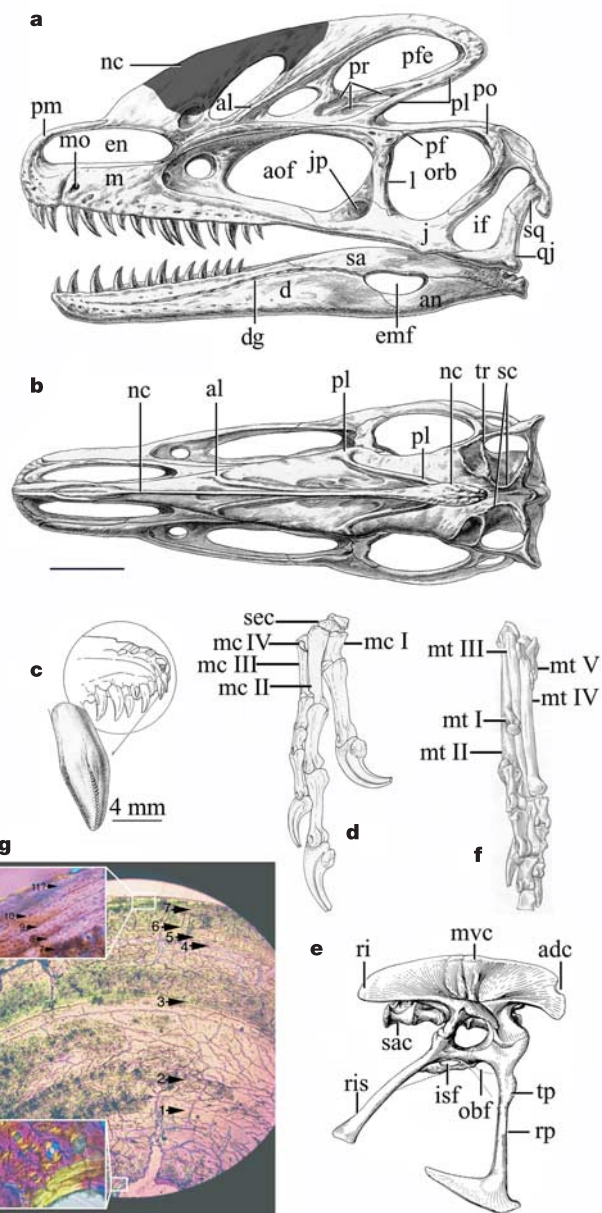


Figure 2 | *Guanlong wucaii* (IVPP V14531). **a**, **b**, Cranial reconstruction in left lateral (**a**; shaded area indicates the unpreserved portion) and dorsal (**b**) views. **c**, Close up of a premaxillary tooth. **d**, Ventral view of left manus. **e**, Pelvis in right lateral view. **f**, Ventral view of left pes. **g**, Histological section from the fibular shaft of *Guanlong wucaii* (IVPP V14531), shown in polarized microscopy (original magnification $\times 6.5$). Numbers and arrows denote growth lines used to age the specimen (see Supplementary Information for a detailed explanation). adc, anterodorsal concavity; al, anterior lamina; an, angular; aof, antorbital fenestra; d, dentary; dg, dentary groove; emf, external mandibular fenestra; en, external naris; if, infratemporal fenestra; isf, foramen on ischium; j, jugal; jp, pneumatic jugal foramen; l, lacrimal; m, maxilla; mc I–IV, metacarpals I–IV; mt I–V, metatarsals I–V; mvc, median vertical crest; nc, nasal crest; obf, obturator foramen; orb, orbit; pf, prefrontal; pfe, pneumatic fenestra; pl, posterior lamina; pm, premaxilla; po, postorbital; pr, pneumatic recess; qj, quadratojugal; ri, right ilium; ris, right ischium; rp, right pubis; sa, surangular; sac, sacrum; sc, sagittal crest; sec, semilunate carpal; sq, squamosal; tp, tubercle on pubis; tr, transverse ridge. Scale bar: 5 cm (**a**, **b**), 7 cm (**d**), 12 cm (**e**) and 8 cm (**f**).

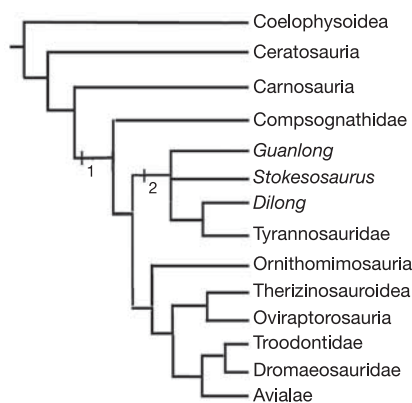


Figure 3 | A simplified cladogram representing an Adams consensus of 6,336 most parsimonious trees showing the phylogenetic position of *Guanlong wucaii* (see Supplementary Information for a detailed analysis). Reduced consensus analysis demonstrates that *Guanlong* is outside of the Tyrannosauridae–*Dilong* clade on all trees, but the position of *Stokesosaurus* is unstable (see Supplementary Information). 1, Coelurosauria; 2, Tyrannosauroidae. Bootstrap value (90%) is high for the Tyrannosauroidae including *Guanlong wucaii*.

vascularization throughout the cortices, broad spacing between growth lines near the periosteal surface, the presence of only a thin, single lamellae of endosteal bone, and just several haversian canals^{20,21} (see Supplementary Information for a more detailed explanation).

Compared to IVPP V14531, IVPP V14532 has a proportionally shorter snout, a shorter pre-antorbital ramus of the maxilla with a less distinct step, a larger orbit, a dorsoventrally taller quadrate, apicobasally shorter teeth, a longer manus relative to the humerus, a longer pubis with a less-developed anterior expansion of the pubic boot, a longer distal segment of the hindlimb, and a slightly posteriorly curved pubic shaft compared to a straight one in IVPP V14531. These features are closer to the conditions present in derived coelurosaurs than are those of IVPP V14531. A few other features, however, are closer to the conditions present in more derived tyrannosauroids than are those of IVPP V14531, such as a shorter humerus relative to the femur, a less bowed ulna, a more prominent anteroventral process of the ilium, and a deeper and less curved post-acetabular process of the ilium.

Guanlong possesses numerous tyrannosauroid apomorphies: large foramina on the lateral surface of the premaxilla; tall premaxillary body; fused nasals; a large frontal contribution to the supratemporal fossa; a pneumatic jugal foramen in the posterior rim of the antorbital fossa; a deep basisphenoidal sinus with large foramina; a subcondylar recess on the basisphenoid; the supraoccipital excluded from the foramen magnum; the short retroarticular process; the relatively small, U-shaped premaxillary teeth that are arranged in a row more transversely than anteroposteriorly oriented; and labiolingually thick maxillary and dentary teeth^{1,15,22,23} (Fig. 2a–c; see also Supplementary Fig. 1). Striking tyrannosauroid pelvic features^{1,2,6} include an ilium subequal to femoral length, a distinctive dorsal concavity on the pre-acetabular process, a supracetabular crest that is straight in dorsal view, a prominent median vertical crest on the lateral surface of the ilium, a concave anterior margin of the pubic peduncle, a pubic tubercle close to the dorsal part of the pubic shaft, an extremely large pubic boot (55% of pubic length), and a thin sheet of bone extending from the obturator process down the ischial shaft (Fig. 2e; see also Supplementary Fig. 2k).

However, *Guanlong* lacks several synapomorphies shared by other tyrannosauroids, such as a reduced external mandibular fenestra, a prominent surangular ridge and a robust manual digit II. Additionally, several features are intermediate between the typical coelurosaurian and tyrannosauroid conditions, such as the relative size of the

premaxillary teeth and the squamosal–quadratojugal flange and the moderately medially inclined ilium. These data suggest a basal-most position for *Guanlong* among the Tyrannosauroidae, as supported by a numerical cladistic analysis (Fig. 3). However, its larger size relative to *Dilong* is inconsistent with the evolutionary trend of increasing size within more derived tyrannosauroids as suggested by previous studies^{11,20}. Furthermore, numerous autapomorphies set *Guanlong* apart from other tyrannosauroids. Thus, *Guanlong* represents a specialized lineage in the early evolution of tyrannosauroids.

The discovery of *Guanlong* also provides an insight into the poorly known early evolution of coelurosaurs. *Guanlong* demonstrates the primitive trait of a pelvic fenestra, also seen in the basal coelurosaur *Mirischia asymmetrica*²⁴ but more typical of basal tetanurans⁵. However, *Guanlong* is similar to derived coelurosaurs in having a large external naris extending posteriorly beyond the anterior margin of the antorbital fossa and a proportionally long forelimb with a bowed ulna and elongate manus⁹. This unusual combination of primitive tetanuran and derived coelurosaurian features suggests a complex pattern of character evolution at the base of the Coelurosauria. Interestingly, *Guanlong* also displays some features that are similar to the proposed synapomorphies of a Therizinosaurioidea–Ornithomimosauria–Alvarezsauridae clade¹⁰, such as a long maxillary pre-antorbital ramus, a proportionally long dentary, a weak surangular ridge and paired flexor processes on the proximal end of manual phalanx II-1 and III-1. It is possible that these features might diagnose a more inclusive group, the Coelurosauria, but were lost in more derived members. Consequently this character distribution shows that the basal conditions for some coelurosaurian clades are highly modified in derived members of each group and underscores the need for more taxonomic sampling at the base of each coelurosaurian group.

The hypertrophied cranial crest of *Guanlong* provides a surprising case for the presence of an exaggerated ornament among non-avian predatory dinosaurs, comparable to well-known exaggerated ornamental traits in other vertebrate groups such as the large antlers of *Megaloceros* (Irish elks) and the long tail of peacocks²⁵. Cranial horns, bosses and crests are present in many non-avian theropods and are best exemplified by *Dilophosaurus*, *Monolophosaurus* and oviraptorids, among others²⁶. The function of these often pneumatic structures in theropods has been considered to be ornamentation involved in display or species recognition²⁷. The cranial crest of *Guanlong*, however, is larger and more elaborate than any yet reported for a non-avian theropod dinosaur²⁶. It seems paradoxical that this predatory taxon possessed a seemingly delicate, highly pneumatized cranial crest. In this regard, *Guanlong*'s cranial crest is similar to the sexually selected ornaments widely present in extant and extinct vertebrates, which have been suggested to exact a viability cost for the bearer²⁵. *Guanlong*'s cranial ornament may be a sexually selected trait, which has also been a suggested explanation for similar structures in some other non-avian dinosaur groups²⁸.

Received 12 October; accepted 5 December 2005.

- Holtz, T. R. in *The Dinosauria* 2nd edn (eds Weishampel, D. B., Dodson, P. & Osmolska, H.) 111–136 (Univ. California Press, Berkeley, 2004).
- Rauhut, O. W. M. A tyrannosauroid dinosaur from the Upper Jurassic of Portugal. *Palaeontology* **46**, 903–910 (2003).
- Eberth, D. A. et al. Sequence stratigraphy, paleoclimate patterns and vertebrate fossil preservation in Jurassic–Cretaceous strata of the Junggar Basin, Xinjiang Autonomous Region, People's Republic China. *Can. J. Earth Sci.* **38**, 1627–1644 (2001).
- Chen, P. J. in *The Continental Jurassic* (ed. Morales, M.) 395–412 (Museum of Northern Arizona, Flagstaff, 1996).
- Holtz, T. R., Molnar, R. E. & Currie, P. J. in *The Dinosauria* 2nd edn (eds Weishampel, D. B., Dodson, P. & Osmolska, H.) 71–110 (Univ. California Press, Berkeley, 2004).
- Rauhut, O. W. M. *The Interrelationships and Evolution of Basal Theropod Dinosaurs. Special Papers in Palaeontology* 69 (Palaeontological Association, London, 2003).
- Gauthier, J. in *The Origin of Birds and the Evolution of Flight* (ed. Padian, K.) 1–55 (California Academy of Sciences, San Francisco, 1986).

8. Norell, M. A., Clark, J. M. & Makovicky, P. J. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. & Gall, L. F.) 49–67 (Yale Univ. Press, New Haven, 2001).
9. Xu, X. *Deinonychosaurian Fossils from the Jehol Group of Western Liaoning and the Coelurosaurian Evolution*. Thesis, Chinese Academy of Sciences (2002).
10. Sereno, P. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. A. & Gall, L. F.) 69–98 (Yale Univ. Press, New Haven, 2001).
11. Xu, X. *et al.* Basal tyrannosauroids from China and evidence for protofeathers in tyrannosauroids. *Nature* **431**, 680–684 (2004).
12. Hutt, S., Naish, D., Martill, D. M., Barker, M. J. & Newbery, P. A preliminary account of a new tyrannosauroid theropod from the Wessex Formation (Early Cretaceous) of southern England. *Cretaceous Res.* **22**, 227–242 (2001).
13. Currie, P. J. New information on the anatomy and relationships of *Dromaeosaurus albertensis* (Dinosauria: Theropoda). *J. Vert. Paleontol.* **15**, 576–591 (1995).
14. Makovicky, P. J. & Norell, M. A. in *The Dinosauria* 2nd edn (eds Weishampel, D. B., Dodson, P. & Osmolska, H.) 184–195 (Univ. California Press, Berkeley, 2004).
15. Brochu, C. A. Osteology of *Tyrannosaurus rex*: insights from a nearly complete skeleton and high-resolution computed tomographic analysis of the skull. *J. Vert. Paleontol. Mem.* **7**, 1–138 (2003).
16. Osborn, H. F. Skull adaptations of *Ornitholestes*, *Struthiomimus*, *Tyrannosaurus*. *Bull. Am. Mus. Nat. Hist.* **35**, 733–771 (1916).
17. Chure, D. J. in *New Perspectives on the Origin and Early Evolution of Birds* (eds Gauthier, J. A. & Gall, L. F.) 283–300 (Yale Univ. Press, New Haven, 2001).
18. Xu, X., Tang, Z.-L. & Wang, X.-L. A therizinosauroid dinosaur with integumentary structures from China. *Nature* **399**, 350–354 (1999).
19. Hinchliffe, J. R. in *The Beginnings of Birds* (eds Hecht, M. K., Ostrom, J. H., Viohl, G. & Wellnhofer, P.) 141–148 (Freunde des Jura-Museum, Eichstatt, 1985).
20. Erickson, G. M. *et al.* Gigantism and comparative life-history parameters of tyrannosaurid dinosaurs. *Nature* **430**, 772–775 (2004).
21. Reid, R. E. H. in *The Complete Dinosaur* (eds Farlow, J. O. & Brett-Surman, M. K.) 403–413 (Indiana Univ. Press, Bloomington, 1997).
22. Currie, P. J. Cranial anatomy of tyrannosaurid dinosaurs from the Late Cretaceous of Alberta, Canada. *Acta Palaeontol. Pol.* **48**, 191–226 (2003).
23. Hurum, J. H. & Sabath, K. Giant theropod dinosaurs from Asia and North America: skulls of *Tarbosaurus bataar* and *Tyrannosaurus rex* compared. *Acta Palaeontol. Pol.* **48**, 161–190 (2003).
24. Naish, D., Martill, D. & Frey, E. Ecology, systematics and biogeographical relationships of dinosaurs, including a new theropod, from the Santana Formation (?Albian, Early Cretaceous) of Brazil. *Hist. Biol.* **16**, 57–70 (2004).
25. Andersson, M. *Sexual Selection* (Princeton Univ., Princeton, 1994).
26. Horner, J. R. Dinosaur reproduction and parenting. *Ann. Rev. Earth Planet. Sci.* **28**, 19–45 (2000).
27. Vickaryous, M. K. & Byan, M. J. in *Encyclopedia of Dinosaurs* (eds Currie, P. J. & Padian, K.) 488–493 (Academic Press, San Diego, 1997).
28. Sampson, S. D., Ryan, M. J. & Tanke, D. H. Craniofacial ontogeny in centrosaurine dinosaurs (Ornithischia: Ceratopsidae): taxonomic and behavioral implications. *Zool. J. Linn. Soc.* **121**, 293–337 (1997).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank D. Naish and T. Holtz for critical comments, H.-J. Wang for organizing the fieldwork, T. Yu for finding the specimen, R. S. Li for illustrations, L. S. Xiang and X.-Q. Ding for preparing the specimen, X.-Q. Ding for editing the illustrations, A. Prieto-Marquez for assistance with the histological figure, M. Carrano for suggestions, and members of the Sino-American expedition team for collecting the fossil. The work in 2002 was supported by the National Geographic Society, the National Natural Science Foundation of China, the Jurassic Foundation, the Hilmar Saltee bequest, George Washington University, and the Chinese Academy of Sciences. Study of the specimens was supported by the National Natural Science Foundation of China, the National Science Foundation Division of Earth Sciences, the Chinese Academy of Sciences, and the American Museum of Natural History.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to X.X. (xingxu@vip.sina.com or xu@amnh.org).

Sympatric speciation in Nicaraguan crater lake cichlid fish

Marta Barluenga^{1*}, Kai N. Stölting^{1*}, Walter Salzburger^{1,2*}, Moritz Muschick¹ & Axel Meyer¹

Sympatric speciation, the formation of species in the absence of geographical barriers, remains one of the most contentious concepts in evolutionary biology. Although speciation under sympatric conditions seems theoretically possible^{1–5}, empirical studies are scarce and only a few credible examples of sympatric speciation exist⁶. Here we present a convincing case of sympatric speciation in the Midas cichlid species complex (*Amphilophus* sp.) in a young and small volcanic crater lake in Nicaragua. Our study includes phylogeographic, population-genetic (based on mitochondrial DNA, microsatellites and amplified fragment length polymorphisms), morphometric and ecological analyses. We find, first, that crater Lake Apoyo was seeded only once by the ancestral high-bodied benthic species *Amphilophus citrinellus*, the most common cichlid species in the area; second, that a new elongated limnetic species (*Amphilophus zaliosus*) evolved in Lake Apoyo from the ancestral species (*A. citrinellus*) within less than ~10,000 yr; third, that the two species in Lake Apoyo are reproductively isolated; and fourth, that the two species are eco-morphologically distinct.

Ever since Darwin proposed the concept in *The Origin*⁷, the issue of whether sympatric speciation is common or even possible in nature has received much attention from evolutionary biologists. Despite current enthusiasm in favour of sympatric speciation and models espousing that it is theoretically possible^{1–5}, the empirical evidence for it remains scant and many of the proposed cases are equally compatible with allopatric schemes. A thorough examination⁶ has led to the conclusion that even the most promising examples of sympatric speciation remain questionable because of the exclusive use of mitochondrial DNA (mtDNA)^{8,9}, small sample sizes^{8,9} or insufficient outgroup sampling¹⁰. It has been emphasized⁶ that a firm corroboration for the origin of one or more species in sympatry would need to fulfil the following criteria: first, a sympatric distribution of the most closely related sister species; second, genetic evidence for reproductive isolation among them; third, their monophyly; and fourth, an ecological setting in which allopatric differentiation is unlikely, such as remote oceanic islands, hosts for parasites or small lakes.

Crater Lake Apoyo in Nicaragua (Fig. 1) provides an exceptionally clear situation for testing sympatric speciation. It is small (diameter, ~5 km; max. depth, ~200 m), is of recent origin¹¹ (<23,000 yr), has a homogenous habitat (Supplementary Fig. 1), and is completely isolated. This filled caldera has an impoverished fauna compared with nearby water bodies¹², but contains the widespread Midas cichlid (*A. citrinellus*, Günther 1864) and an endemic species, the Arrow cichlid (*A. zaliosus*, Barlow 1976). These, together with *Amphilophus labiatus* from the large Nicaraguan lakes^{13–15}, and possibly more species from some of the crater lakes¹⁶, make up the Midas cichlid species complex that is restricted to Nicaragua and the

north of Costa Rica. In *A. citrinellus* of other lakes and in *A. labiatus*, two colour morphs are recognized, but such polymorphism has not been described in Lake Apoyo. To examine whether the endemic Arrow cichlid emerged in Lake Apoyo from an ancestral stock of Midas cichlids under fully sympatric circumstances, we adopted a comprehensive approach including phylogeographic, population-genetic, morphometric and ecological analyses. We included about 120 specimens from Lake Apoyo and over 500 individuals from six other lakes of Nicaragua (outgroup sample; see Supplementary Table 1).

Analysis of the mitochondrial control region (840 bp) showed that the two *Amphilophus* species from Lake Apoyo form a monophyletic assemblage (Fig. 2a). Notably, not a single mtDNA haplotype is

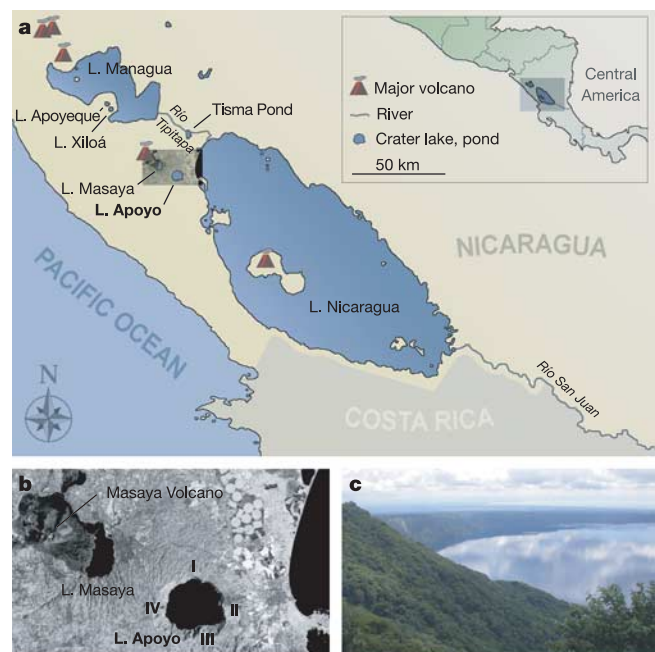


Figure 1 | The study area. **a**, Map of the Pacific coast of Nicaragua and Costa Rica showing the large Nicaraguan lakes (Managua, Nicaragua), some of the volcanic crater lakes (Apoyo, Apoyeque, Masaya and Xiloá), Tisma Pond and the rivers Tipitapa and San Juan that are home to the Midas cichlid species complex (*Amphilophus* sp.). **b**, Satellite image showing Lake Apoyo and neighbouring Lake Masaya that has partially been overflowed by lava from a more recent eruption of the Masaya Volcano. Roman numerals indicate the four sampling localities. **c**, Photograph of Lake Apoyo showing its steep crater wall.

¹Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, and ²Center for Junior Research Fellows, University of Konstanz, Universitätsstrasse 10, 78457 Konstanz, Germany.

*These authors contributed equally to this work.

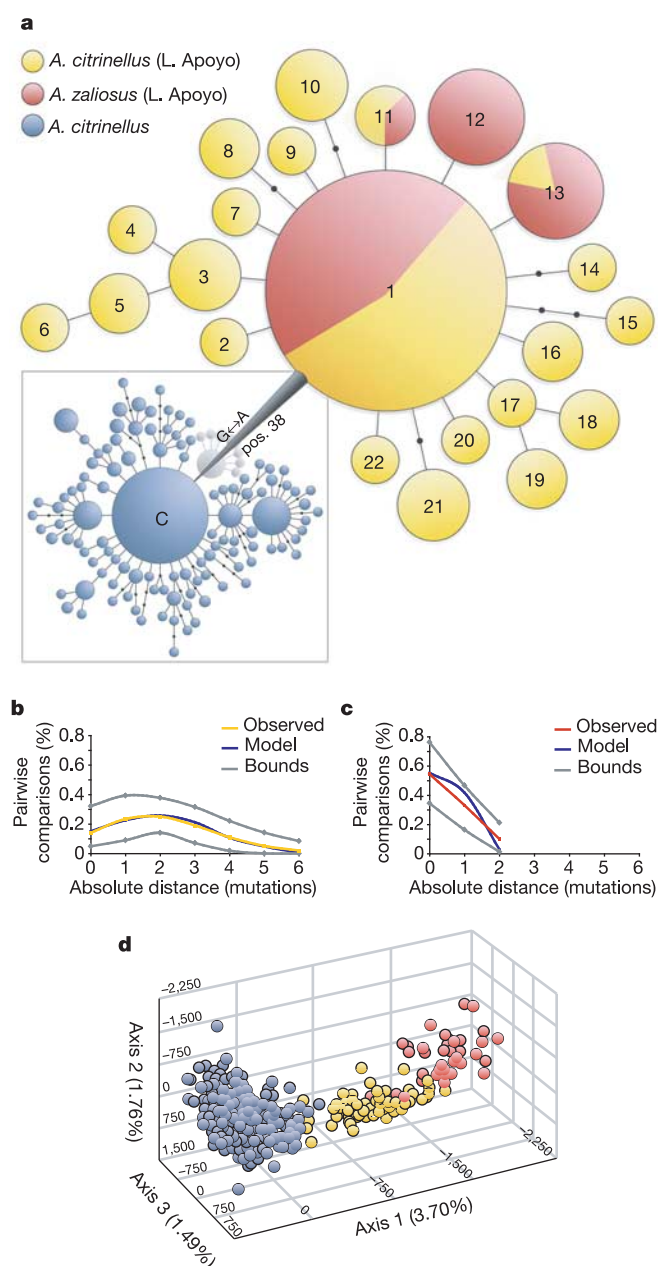


Figure 2 | Phylogeography of the Lake Apoyo *Amphilophus* species.

a, Unrooted haplotype network based on 637 mtDNA sequences. Haplotypes are coloured according to species and geographic origin: *A. zaliosus* (red), *A. citrinellus* from Lake Apoyo (yellow), and *A. citrinellus* from other lakes (blue). The framed haplotype network is taken from ref. 15 and is based on 519 specimens from all lakes except Lake Apoyeque (the 30 Lake Apoyo specimens included in that study are shown in grey). The new Lake Apoyo haplotypes are enlarged; their sizes reflect the number of specimens sharing the same haplotype (for example: haplotype 1, 65 individuals; haplotype 2, 1 individual). Lake Apoyo representatives, albeit belonging to two species, form a monophyletic assemblage showing a characteristic mutation in alignment position 38. The central haplotype in Lake Apoyo (1) is connected by this characteristic mutation to the central haplotype C. **b**, Coalescence-based mitochondrial mismatch analysis of Lake Apoyo *A. citrinellus* uncovers a demographic expansion about two mutations ago. **c**, Mismatch analysis of *A. zaliosus* suggests ongoing demographic expansion (see Supplementary Table 2 for parameters). **d**, Three-dimensional representation of a factorial correspondence analysis based on microsatellite genotypes. Individuals are colour-coded as described above.

shared between Lake Apoyo and any of the remaining lakes, suggesting that there was a single colonization event by an ancestral lineage of *A. citrinellus* and ruling out any contemporary genetic exchange or secondary colonizations. The genetic diversity and the number of mtDNA haplotypes are low in Lake Apoyo (22 versus 125 in the outgroup sample), in line with the young age of the species assemblage¹⁵. A mitochondrial mismatch analysis indicates a single demographic expansion of *A. citrinellus* in Lake Apoyo about two mutations ago (Fig. 2b), whereas a more recent demographic expansion is detected for the new species *A. zaliosus* (Fig. 2c and Supplementary Table 2). A plot of the microsatellite distances (Fig. 2d and Supplementary Fig. 2 and Table 3) and the phylogenies based on microsatellites and amplified fragment length polymorphisms (AFLPs; Supplementary Fig. 3) corroborate the distinctive genetic composition of Lake Apoyo's *Amphilophus* fauna, its monophyly and the evolution of *A. zaliosus* from the *A. citrinellus* stock of Lake Apoyo.

Significant *F*-statistics based on all genetic markers show that *A. citrinellus* and *A. zaliosus* are reproductively isolated in Lake Apoyo (Table 1). Similarly, microsatellite-based bayesian population assignment tests (Fig. 3) unambiguously identified three clusters: the outgroup sample, *A. citrinellus* and *A. zaliosus* from Lake Apoyo. Reproductive isolation is further suggested by mate-choice experiments that demonstrate strong assortative mating between both species¹⁷.

The two *Amphilophus* species in Lake Apoyo are morphologically clearly distinct^{13,14} (Fig. 4a). Morphometric analysis uncovered two discrete body types corresponding to the two species, with body height explaining most of the differences (Fig. 4b). The two species also differ in the shape of a trophically relevant structure that is tightly linked to the ability of cichlids to process alternative food types¹⁸: the pharyngeal jaw (Fig. 4c). Morphometric analyses show clear differences between the two species (Fig. 4d), with *A. zaliosus* having more elongated pharyngeal jaws. Stomach content analyses indicate significant dietary differences between the two species and a wider trophic niche in *A. citrinellus* (Fig. 4e). Our eco-morphological inferences characterize *A. citrinellus* as benthic forager with a deeper body, and *A. zaliosus* as limnetic form with an elongated trunk that appears to be adapted to living in the open water column.

Taken together, we present a strong case where only sympatric speciation can account for the origin of a new species from a more widespread one in Lake Apoyo in <10,000 yr (Supplementary Table 2). The Lake Apoyo population of the Midas cichlid and the Arrow cichlid form a monophyletic assemblage (Fig. 2 and Supplementary Fig. 3); they are reproductively completely isolated as shown by mate-choice experiments¹⁷ and the analyses of three sets of molecular markers (Table 1); they are sympatrically distributed, no genetic structuring was detected in *A. citrinellus* and *A. zaliosus*, and even individuals from opposite sides of the crater lake show no sign of differentiation (Supplementary Table 4, Supplementary Fig. 4). The recent volcanic origin of Lake Apoyo, its small size, its degree of isolation, the homogeneous habitat and the sympatric occurrence of both species throughout the lake, as well as the absence of genetic structure in each of the two species rule out the possibility of (micro-) allopatric or parapatric differentiation.

Disruptive natural selection is likely to explain the evolution of alternative habitat preferences in Lake Apoyo's cichlid species.

Table 1 | Genetic distance between the two *Amphilophus* species from Lake Apoyo

<i>F</i> -statistics	mtDNA (840 bp)	Microsatellites (10 loci)	AFLPs (226 variable loci)
<i>A. citrinellus</i> versus <i>A. zaliosus</i>	0.06**	0.12**	0.06**

The genetic distance was measured with *F*-statistics (***P* < 0.001).

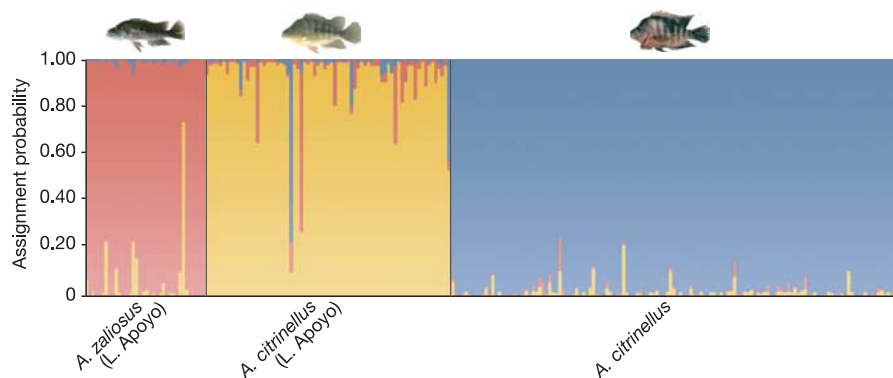


Figure 3 | Bayesian population assignment test. A population assignment test with the software Structure based on ten microsatellite loci uncovered

However, most theoretical models predict that the completion of speciation requires the evolution of other forms of pre- or post-zygotic isolation⁵, such as assortative mating, and a linkage disequilibrium between loci involved in disruptive natural selection and those involved in mate choice^{1,3}. Both field observations¹³ and laboratory experiments¹⁷ show that the two *Amphilophus* species in Lake Apoyo mate assortatively. Although heterospecific pairs can form, they do not lead to successful matings¹⁷, indicating that differences in courtship behaviour result in effective prezygotic isolation. Therefore, both ecological speciation mechanisms through divergent habitat preferences and resource partitioning, and assortative mating through behavioural isolation separate the two *Amphilophus* species in Lake Apoyo.

The finding that the two *Amphilophus* species from Lake Apoyo differ in eco-morphological traits and resource utilization (Fig. 4) indicates that ecological disruptive selection played a principal part in the speciation event. Similar processes of divergent niche occupation are also known from other species of freshwater fishes¹⁹. For Lake Apoyo, it seems likely that a new ecological dimension—the open water column over greater depths, which is absent in the shallow large Nicaraguan lakes—facilitated the emergence of a limnetic morphotype and, ultimately, of the new species *A. zaliosus*. In this respect it is interesting that in the second deepest lake in Nicaragua, Crater Lake Xiloá (maximum estimated depth, 100 m), an elongated morphotype of the Midas cichlid occurs that also seems to be adapted to the open water column, but is not related to *A. zaliosus*^{15,16}. Although details about ecology, morphology and genetic structure of Lake Xiloá's elongated morphotype remain tentative, it seems plausible that the Xiloá situation parallels the one in Lake Apoyo, rendering the Midas cichlid assemblage as another example of parallel sympatric speciation in addition to the sticklebacks in postglacial lakes in British Columbia¹⁹.

The outstanding species richness of cichlids in the East African Great Lakes, which alone contain more than 1,500 endemic species^{20,21}, has repeatedly been suggested to be caused by sympatric speciation through sexual selection, as most cichlid species differ more strongly in body coloration than in morphology^{22,23}. However, the enormous dimensions of these lakes, their habitat complexity, and the complicity to infer sister species relationships among closely related taxa in those systems make it difficult, if not impossible, to discard allopatric or parapatric schemes. But one of the few possible cases of sympatric speciation involves cichlids, namely tilapiine cichlids in crater lakes in Cameroon^{6,8,24}, where, as in Lake Apoyo, disruptive natural selection seems to be the driving force of the formation of new species. This is in line with data showing that sympatric speciation by sexual selection alone is rather unlikely^{25,26}. The finding that sympatric speciation accounts for the origin of at least one new species in the Midas cichlids, and probably also for the emergence of new species in West African crater lakes, again under-

three distinct populations (*A. zaliosus*, red; *A. citrinellus* from Lake Apoyo, yellow; and *A. citrinellus* from Lake Nicaragua, blue).

scores the special propensity of cichlids for sympatric speciation. Further empirical studies, new theoretical approaches and, in particular, the synthesis of empirical data and simulations are necessary to justify the extrapolation from a small model system such as Lake Apoyo to the exceptionally diverse cichlid species flocks of East African lakes Victoria, Malawi and Tanganyika.

METHODS

Sampling. A detailed list of all specimens included in this study, voucher identifications, sampling localities, mtDNA haplotypes and GenBank accession numbers is provided in Supplementary Table 1. Fish were collected in the fall of 2003 with gill nets and photographed. Fin clips preserved in 90% ethanol were taken as DNA samples. Both Lake Apoyo species co-occurred at all sampling sites (Fig. 1 and Supplementary Table 1). Fish heads were taken as voucher specimens and to dissect the lower pharyngeal jaws. In addition, full stomachs of some sampled individuals were taken for dietary analyses. Sampling was done with permit 033-0903 of the Nicaraguan Ministry of Natural Resources (MARENA).

Stomach content analysis. The stomachs were dissected using a dissecting stereomicroscope, and the contents were classified into five categories: biofilm (amorphous blue-green substance mixed with sand¹²), the algae *Chara* and other plant material, zooplankton, insects and fish remnants. The volume of each category present in each individual was estimated and used to calculate Schoener's index of diet overlap²⁷ (values greater than 0.6 generally indicate biologically significant overlap) and trophic niche breadth²⁸ of both species.

Morphometric analyses. Nine landmarks describing the body shape of the fish from Lake Apoyo (81 *A. citrinellus*, 35 *A. zaliosus*) and ten landmarks describing the shape of their pharyngeal jaws (30 *A. citrinellus*, 18 *A. zaliosus*) were digitized from images taken with a Nikon Coolpix 995 digital camera using TpsDig2 (ref. 29). Two-dimensional Procrustes fitting was done with Past³⁰ to standardize landmark coordinates, followed by a shape principal component analysis (PCA) to identify the variables that accounted for the largest variance in the data. Species were discriminated on the basis of established meristic measurements¹³.

mtDNA analyses. Total DNA was extracted by standard salt extraction and the whole control region was amplified by using published primers and PCR conditions¹⁵. A haplotype network was constructed from the maximum-likelihood tree constructed with Paup* 4.0b10 (<http://www.sinauer.com>) that was translated into maximum parsimony branch lengths. A coalescence-based mismatch analysis was done with Arlequin 2.1 (<http://lgb.unige.ch/arlequin>) to study the demographic history and to estimate the relative timing of the population expansion of the fish in Lake Apoyo.

Microsatellite genotyping. Individuals were genotyped with ten unlinked microsatellite loci (those in ref. 15, and Unh011, Unh012 and Unh013; ref. 16), amplified with fluorescently labelled reverse primers (HEX and FAM dyes), and analysed on an ABI 3100 automatic capillary sequencer. Fragment visualization and scoring were done as previously described¹⁵.

AFLP amplification. AFLP analysis was done by standard protocols using a plant mapping kit (Applied Biosystems). Five primer combinations were used for selective amplification (*EcoRI* + *MseI* + : ACA/CAA, ACA/CTG, ACC/CAA, ACC/CAG, AGC/CAA); the *EcoRI* primers were fluorescently labelled (FAM). Visualization and scoring of fragments were done as described for microsatellites. The presence or absence of fragments was binary coded (1 or 0),

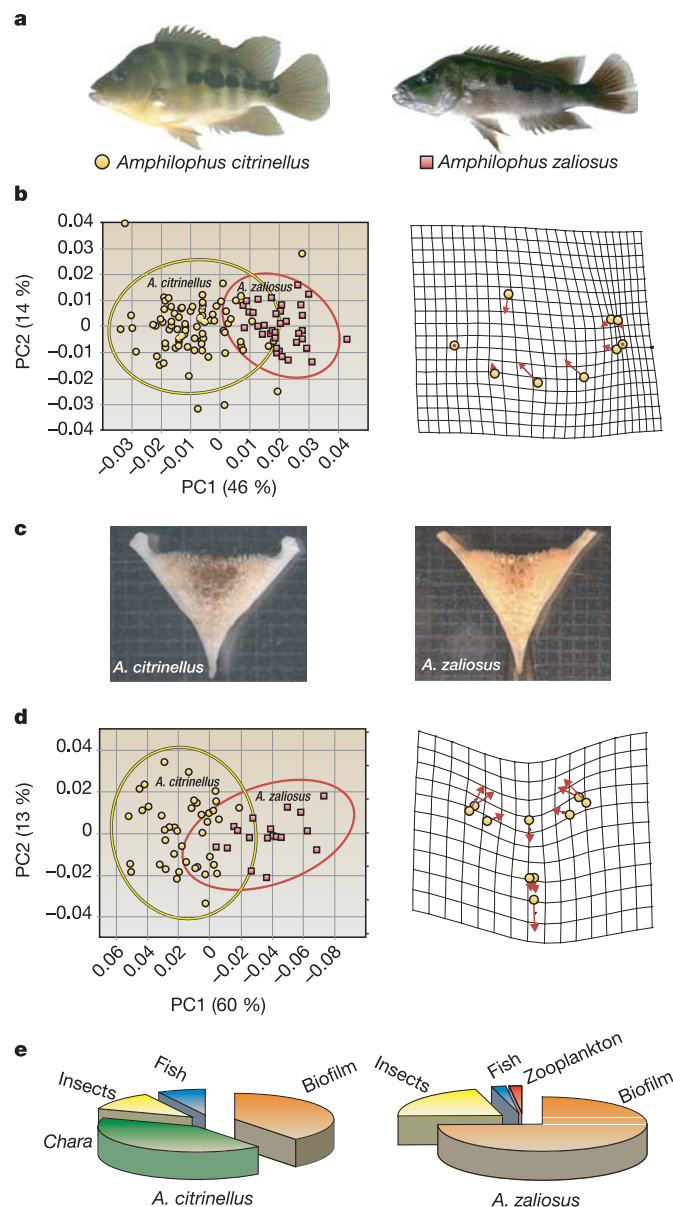


Figure 4 | Eco-morphological assessment of Lake Apoyo's

Amphilophus sp. **a**, The Midas cichlid (*Amphilophus citrinellus*) and the Arrow cichlid (*A. zaliosus*) are morphologically distinct. **b**, A PCA based on nine landmarks (left) shows a clear separation in body shape between *A. citrinellus* and *A. zaliosus* with little overlap. Ellipses represent the 95% confidence intervals. The transformation grid based on principal component 1 (right; scaling factor 2) indicates that body height accounts for most of the variance between the two species. **c**, *A. citrinellus* and *A. zaliosus* show clear differences in their pharyngeal jaws. **d**, A PCA based on ten landmarks (left) shows clear differences in this structure between the two species. The corresponding transformation grid (right; scaling factor 2) indicates that the pharyngeal jaws of *A. zaliosus* are more elongated. In addition, *A. zaliosus* has only relatively small pointed teeth, whereas some *A. citrinellus* have molariform teeth. **e**, Stomach content analyses reveal clear-cut differences between the two Lake Apoyo species (Schoener index: 0.53). Both species forage on biofilm, but whereas winged insects were found in all individuals of *A. zaliosus*, insects were recovered in only about half of the *A. citrinellus* stomachs. Instead, the latter feed on the algae *Chara* and other plant material that is found close to the lake's shore¹². The feeding niche of *A. citrinellus* is wider than that of *A. zaliosus* (Levins index: $B_{\text{citrinellus}} = 3.01$, $B_{\text{zaliosus}} = 1.85$).

and a distance matrix was constructed with Arlequin for statistical analyses.

Population genetic parameters and phylogenetic reconstruction. Allelic diversity, genetic variation, deviation from Hardy–Weinberg equilibrium, and genetic differentiation were calculated with Arlequin. *F*-statistics were calculated for all molecular markers as implemented in Arlequin. Significance levels were Bonferroni-corrected. Using microsatellite data, we applied a bayesian model-based clustering algorithm implemented in Structure 2.1 (<http://pritch.bsd.uchicago.edu>) to test the assignment of the cichlids from Lake Apoyo. We used the admixture model and determined the number of ancestral clusters, *K*, by comparing log-likelihood ratios in multiple runs for values of *K* between 1 and 5. Each run consisted of 1,000,000 iterations with a burn-in period of 50,000; multiple runs with the same value of *K* led to virtually identical results. To detect the degree of similarity of the Midas cichlids in Lake Apoyo to those outside the lake, a factorial correspondence analysis of the microsatellite data was done with Genetix 4.0 (<http://www.univ-montp2.fr/~genetix/genetix/genetix.htm>), which graphically projects the individuals on the factor space defined by the similarity of their allelic states. We inferred a phylogeny of different populations of Midas cichlids on the basis of the microsatellite data with Contml of the Phylip package (<http://evolution.genetics.washington.edu/phylip.html>), applying a brownian motion approximation using the mean allelic copy number, and with Populations (<http://www.cnrs-gif.fr/pge/bioinfo/populations/index.php>) using the Cavalli–Sforza and Edwards method (10,000 bootstrap replicates on locus information). For both analyses individuals were grouped according to species and lake of origin. We used Paup* to reconstruct a phylogeny based on AFLP data (neighbour-joining algorithm; Nei–Li restriction site differences).

To test for the absence of genetic structuring in Lake Apoyo species, we calculated pairwise *F*-statistics among fish from the four sampling localities with Arlequin (we omitted the data of site I of *A. zaliosus* owing to the small sample size). In addition, we did a bayesian analysis of population structure with Structure for each species. We also ran Migrate (<http://popgen.csit.fsu.edu/>) to estimate the number of migrants between sample sites within each species.

Received 13 September; accepted 10 October 2005.

- Dieckmann, U. & Doebeli, M. On the origin of species by sympatric speciation. *Nature* **400**, 354–357 (1999).
- Higashi, M., Takimoto, G. & Yamamura, N. Sympatric speciation by sexual selection. *Nature* **402**, 523–526 (1999).
- Kondrashov, A. S. & Kondrashov, F. A. Interactions among quantitative traits in the course of sympatric speciation. *Nature* **400**, 351–354 (1999).
- Gavrilets, S. *Fitness Landscapes and the Origin of Species: Monographs in Population Biology* vol. 41 (Princeton Univ. Press, Princeton, NJ, 2004).
- Kawecki, T. J. in *Adaptive Speciation* (Dieckmann, Doebeli, Metz & Tautz, Cambridge Univ. Press, Cambridge, UK, 2004).
- Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer Associates, Sunderland, MA, 2004).
- Darwin, C. *On the Origin of Species by Means of Natural Selection or the Preservation of Favoured Races in the Struggle for Life* (J. Murray, London, 1859).
- Schliwen, U., Tautz, D. & Pääbo, S. Sympatric speciation suggested by monophyly of crater lake cichlids. *Nature* **368**, 629–632 (1994).
- Weiblen, J. R. & Bush, G. L. Speciation in fig pollinators and parasites. *Mol. Ecol.* **11**, 1573–1578 (2002).
- Gíslason, D., Ferguson, M. M., Skúlason, S. & Snorasson, S. S. Rapid and coupled phenotypic differentiation in Icelandic Arctic charr (*Salvelinus alpinus*). *Can. J. Fish Aquat. Sci.* **56**, 2229–2234 (1999).
- Bice, D. Quaternary volcanic stratigraphy of Managua, Nicaragua: Correlation and source assignment for multiple overlapping plinian deposits. *Geol. Soc. Am. Bull.* **96**, 553–566 (1985).
- Barlow, G. W. in *Investigations of the Ichthyofauna of Nicaraguan Lakes* (ed. Thorson, T. B.) 333–358 (Univ. Nebraska Press, Nebraska, 1976).
- Barlow, G. W. & Munsey, J. W. in *Investigations of the Ichthyofauna of Nicaraguan Lakes* (ed. Thorson, T. B.) 359–369 (Univ. Nebraska Press, Nebraska, 1976).
- Klingenberg, C. P., Barluenga, M. & Meyer, A. Body shape variation in cichlid fishes of the *Amphilophus citrinellus* species complex. *Biol. J. Linn. Soc.* **80**, 397–408 (2003).
- Barluenga, M. & Meyer, A. The Midas Cichlid species complex: incipient sympatric speciation in Nicaraguan cichlid fishes? *Mol. Ecol.* **13**, 2061–2076 (2004).
- McKaye, K. R. et al. Behavioral, morphological and genetic evidence of divergence of the Midas cichlid species complex in two Nicaraguan crater lakes. *Cuadernos de Investigación de la UCA* **12**, 19–47 (2002).
- Baylis, J. R. A quantitative study of long-term courtship: I. Ethological isolation between sympatric populations of the midas cichlid, *Cichlasoma citrinellum*, and the arrow cichlid, *C. zaliosum*. *Behaviour* **59**, 59–69 (1976).

18. Meyer, A. Costs and benefits of morphological specialization: feeding performance in the trophically polymorphic Neotropical cichlid fish, *Cichlasoma citrinellum*. *Oecologia* **80**, 431–436 (1989).
19. Schluter, D. & McPhail, J. D. Character displacement and replicate adaptive radiation. *Trends Ecol. Evol.* **8**, 197–200 (1993).
20. Fryer, G. & Iles, T. D. *The Cichlid Fishes of the Great Lakes of Africa. Their Biology and Evolution* (Oliver & Boyd, Edinburgh, 1972).
21. Kocher, T. D. Adaptive evolution and explosive speciation: the cichlid model. *Nature Rev. Genet.* **5**, 288–298 (2004).
22. van Oppen, M. J. H. *et al.* Assortative mating among rock-dwelling cichlid fishes supports high estimates of species richness for Lake Malawi. *Mol. Ecol.* **7**, 991–1001 (1998).
23. Seehausen, O. & van Alphen, J. J. M. Can sympatric speciation by disruptive sexual selection explain rapid evolution of cichlid diversity in Lake Victoria? *Ecol. Lett.* **2**, 262–271 (1999).
24. Schlieffen, U. *et al.* Genetic and ecological divergence of a monophyletic cichlid species pair under fully sympatric conditions in Lake Ejagham, Cameroon. *Mol. Ecol.* **10**, 1471–1488 (2001).
25. Arnegard, M. E. & Kondrashov, A. S. Sympatric speciation by sexual selection alone is unlikely. *Evolution* **58**, 222–237 (2004).
26. Kirkpatrick, M. & Nuismer, S. L. Sexual selection can constrain sympatric speciation. *Proc. R. Soc. Lond. B.* **271**, 687–693 (2004).
27. Schoener, T. W. Non-synchronous spatial overlap of lizards in patchy habitats. *Ecology* **60**, 703–710 (1970).
28. Levins, R. *Evolution in Changing Environments. Some Theoretical Explorations* (Princeton Univ. Press, Princeton, NJ, 1968).
29. Rohlf, F. J. TpsDig. (Dept. Ecology and Evolution, SUNY, New York, 2004).
30. Hammer, Ø., Harper, D. A. T. & Ryan, P. D. Past: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica* **4**, art. 4 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Ministry of Natural Resources (MARENA) in Nicaragua for collection permits; Z. Pérez for support in the field; C. Chang-Rudolf, K. Hofmann, Y. Chiari and N. Feiner for technical assistance in the laboratory; and S. Gavrillets, T. Price and E. B. Taylor for comments on the manuscript. This work was funded by the Deutsche Forschungsgemeinschaft priority program 1127 ('Adaptive radiations').

Author Information GenBank accession numbers for all mitochondrial control region sequences can be found in Supplementary Table 1. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.M. (Axel.Meyer@uni-konstanz.de).

LETTERS

Multiplex amplification of the mammoth mitochondrial genome and the evolution of Elephantidae

Johannes Krause¹, Paul H. Dear², Joshua L. Pollack³, Montgomery Slatkin³, Helen Spriggs², Ian Barnes⁴, Adrian M. Lister⁴, Ingo Ebersberger⁵, Svante Pääbo¹ & Michael Hofreiter¹

In studying the genomes of extinct species, two principal limitations are typically the small quantities of endogenous ancient DNA and its degraded condition¹, even though products of up to 1,600 base pairs (bp) have been amplified in rare cases². Using small overlapping polymerase chain reaction products, longer stretches of sequences or even whole mitochondrial genomes^{3,4} can be reconstructed, but this approach is limited by the number of amplifications that can be performed from rare samples. Thus, even from well-studied Pleistocene species such as mammoths, ground sloths and cave bears, no DNA sequences of more than about 1,000 bp have been reconstructed^{5–7}. Here we report the complete mitochondrial genome sequence of the Pleistocene woolly mammoth *Mammuthus primigenius*. We used about 200 mg of bone and a new approach that allows the simultaneous retrieval of multiple sequences from small amounts of degraded DNA. Our phylogenetic analyses show that the mammoth was more closely related to the Asian than to the African elephant. However, the divergence of mammoth, African and Asian elephants occurred over a short time, corresponding to only about 7% of the total length of the phylogenetic tree for the three evolutionary lineages.

We have developed a multiplex polymerase chain reaction (PCR) approach that in principle allows an entire mitochondrial genome to be amplified from ancient DNA using just two initial amplifications. This is accomplished by using primer pairs that define overlapping DNA sequence fragments representing the complete mitochondrial genome. These primer pairs are combined into two sets, each containing every second primer pair. Each of these two sets is used in a multiplex PCR amplification that requires the same amount of ancient template DNA as would be used for amplifying a single target sequence. Subsequently, the two primary amplifications are diluted and used as templates in secondary PCR reactions, in which each product is amplified individually.

To test whether this approach works for Pleistocene DNA, we designed 46 primer pairs, which together we expected to amplify the entire mitochondrial (mt) genome of the woolly mammoth (Fig. 1; see Supplementary Information). Forty-one of the secondary amplifications yielded products of the expected size (Fig. 2 and Supplementary Information). Of the five PCR amplifications that failed, four were successful in later attempts, suggesting that the initial failure was due to the absence of even a single template molecule in the PCR amplification. The fifth PCR failed repeatedly, and inspection of the sequences of the overlapping amplicons revealed differences at the priming sites between the mammoth and elephant sequences that can account for the failure. New primers were

designed for this amplicon on the basis of the flanking mammoth sequence. The sequences of all 46 amplicons were found to be similar to (but distinct from) corresponding sequences of extant elephants.

To ensure the authenticity of the ancient DNA sequences¹, we repeated the primary and secondary amplifications and the sequencing of the amplicons (see Methods). In 17 of the amplicons, we found differences at 1–5 nucleotide positions between the two independent experiments, presumably due to cytosine deamination of the template⁸ (Supplementary Information). In these cases, a third round of primary PCR amplifications was performed, and the consensus of the three sets of sequences was inferred to represent the correct sequence. In total, fourteen primary PCR amplifications, corresponding to ~200 mg of mammoth bone, were required for determination and confirmation of the entire mtDNA sequence.

To further confirm the reproducibility of the results, samples of the bone were extracted, amplified and sequenced in laboratories in Cambridge and London, UK, that did not have access to the initial results obtained in Leipzig. In total, 5,024 bp of the mitochondrial sequence were independently reproduced. The reproduced sequences

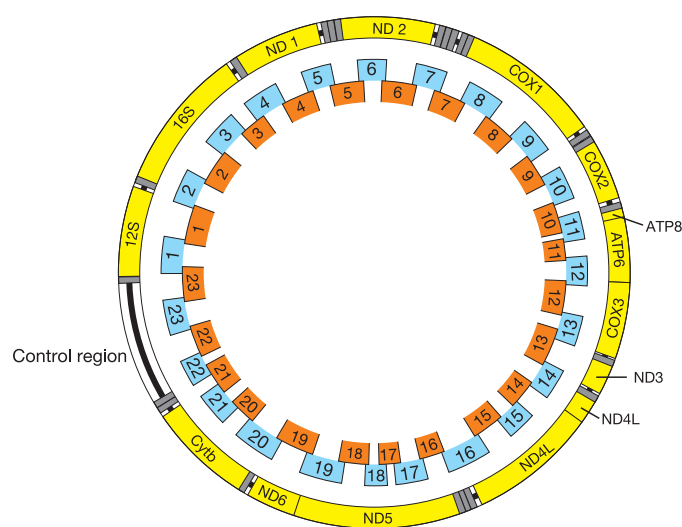


Figure 1 | Map of the mammoth mitochondrial genome. Circular genome (yellow), showing the positions of the control region and the genes encoding 22 transfer RNAs (grey boxes), 2 ribosomal RNAs and 13 proteins. The positions and relative lengths of the 46 amplification products used are depicted in blue (first set) and red (second set).

¹Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany. ²MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK. ³Department of Integrative Biology, University of California, Berkeley, California 94720-3140, USA. ⁴Department of Biology, University College London, London WC1E 6BT, UK. ⁵WE Informatik, Heinrich Heine Universität, Universitätsstr. 1, D-40225 Düsseldorf, Germany.

were identical to those initially determined, except for one position determined from a single PCR reaction in London that shows a thymine (T), when the sequence determined from several PCR amplifications in Leipzig shows a cytosine (C). This discrepancy is most likely due to cytosine deamination of the template DNA⁸.

One concern when determining mtDNA sequences is that nuclear insertions of mtDNA fragments might be mistaken for the organelle copies⁹. This problem can be particularly severe in some species, including elephants¹⁰. However, the sequence we have determined is circular and therefore cannot represent a nuclear insertion. Although it is conceivable that some of the amplimers could have arisen from nuclear insertions, no mismatches were observed in a total of 2,030 overlapping base pairs between adjacent amplimers. We thus conclude that the results obtained indeed represent the complete mtDNA sequence of the mammoth from Berelekh.

The phylogenetic relationship of the mammoth to its closest living relatives, the African and Asian elephants, has until now been unresolved. Numerous studies using small mitochondrial and nuclear DNA¹¹ sequences have variously suggested that mammoths were the sister group of African^{12–14} or of Asian elephants^{7,11,15}. It has been argued that the failure to resolve this phylogenetic relationship is due to the small amount of DNA sequence information available¹³, because long mtDNA sequences are often necessary to obtain a robust phylogenetic tree¹⁶. To address this question, we aligned the entire 16,770-bp mammoth mtDNA with mtDNA from the two elephant species and two potential outgroup species, dugong and hyrax. Phylogenetic trees identified either the Asian or African elephant as the sister taxon of the mammoth, depending on the outgroup and tree reconstruction method used (Supplementary Table S2).

Two lines of evidence indicate that this lack of resolution results from a too-distant relationship of elephants and mammoths to the two outgroup species, both of which diverged from the proboscidean lineage at least 65 million years ago¹⁷. First, about half of the

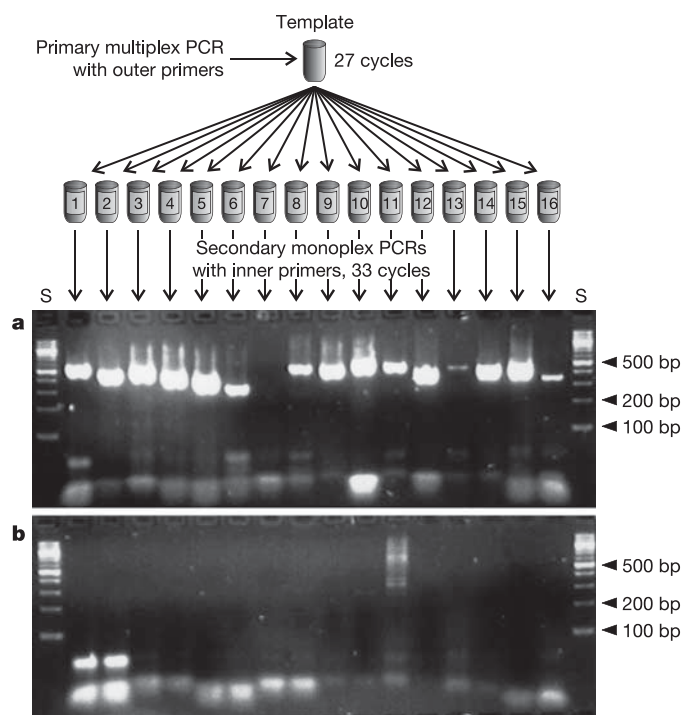


Figure 2 | Principle of the multiplex approach and typical results.

Amplification results for the mammoth ancient DNA extract, with 15 of 16 amplifications yielding positive results (a). Negative controls for the same 16 primer pairs are shown in b. Gel lanes marked with S contain molecular size markers. The size of the amplification products ranges from 306 to 554 bp.

third-codon positions in the protein-coding sequences differ when the mammoth or either elephant species is compared to dugong or hyrax. Second, the ratio of transitions to transversions among elephants and mammoth is 20:1, but is 2:1 between these species and dugong or hyrax. Thus, the phylogenetic signal is blurred by multiple substitutions when dugong or hyrax is included in the analysis¹⁸.

As no closer outgroup species exists, we restricted our analysis to the mammoth and the two elephant species using three different methods that assume a molecular clock (that is, a uniform rate of nucleotide substitutions in all lineages). We first tested whether the mtDNAs of the three species evolved at equal rates using a clock test that uses only three taxa and treats transitions and transversions separately¹⁹, as well as by a likelihood ratio test. The assumption of a molecular clock among the mammoth and two elephant mtDNAs could not be rejected ($P = 0.61$). We therefore estimated a maximum-likelihood tree under the assumption of a molecular clock, applying midpoint rooting. This analysis supports a sister-group relationship between mammoth and Asian elephant with a 97% bootstrap value. We then calculated the likelihood of the data assuming a star-like tree topology. A likelihood ratio test reveals that the more complex model, resulting in the resolved sequence tree, explains the data significantly better, and thus rejects the simpler model of a star-like tree ($P < 0.01$). Notably, the tree that places the Asian elephant and mammoth as the closest relatives to each other has a posterior probability of 99.8%. This indicates that the two alternative scenarios, each having posterior probabilities of $\sim 0.1\%$, can be neglected. Finally, we used Felsenstein's²⁰ method for testing whether the most parsimonious tree is significantly different from a tree with a 'trifurcation' (Fig. 3a). Again, we rejected the null hypothesis of a trifurcation ($P = 0.032$ and $P = 0.035$ for the C and S statistics, respectively; see Methods). Thus, the monophyly of the mammoth and the Asian elephant, to the exclusion of the African elephant, is significantly supported.

In this context, two points should be noted. First, the length of the internal branch leading to the mammoth and Asian elephant mtDNA is only 7.3% of that leading to the African elephant (Fig. 3b). Paleontological data²¹ suggest a divergence of Asian and African elephants and mammoths about six million years ago in Africa. This

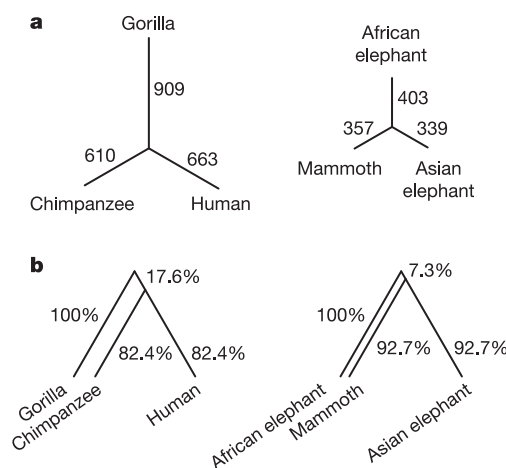


Figure 3 | Comparison of three-species trees. Comparisons based on the complete mtDNA sequences for human, chimpanzee, gorilla, mammoth and African and Asian elephants. **a**, Absolute number of phylogenetically informative positions on which the three-species test is based. **b**, Relative length of the internal branch in the two groups of species for clock-like phylogenetic trees using the substitution data from **a** and rooted by midpoint rooting. Note the differences in the absolute numbers of substitutions and relative branch length of the internal branch between the two species groups (see also Supplementary Information).

date implies that the divergence between the mammoth and Asian elephant took place only 440,000 years after the divergence of the African elephant. Second, this time is short enough that polymorphisms in the ancestral species may have persisted between the two speciation events²², as has been observed for humans, chimpanzees and gorillas²³. As the probability of such events increases with increasing effective population size²⁴, mtDNA is more likely to reflect species phylogeny than autosomal genes that have an approximately four times larger effective population size. The sister-group relationship of the mtDNAs of the mammoth and the Asian elephant therefore suggests that these two species shared a common ancestor after their separation from the African elephant (Supplementary Information). Thus, any morphological similarities between the African elephant and mammoth¹³ are likely to be either ancestral or convergent. However, owing to the short time between the two divergence events, this is also likely to be true for many morphological similarities between the Asian elephant and mammoth.

In summary, the multiplex approach described here allows the retrieval of complete mtDNA sequences from Late Pleistocene fossils, enabling robust phylogenetic inferences to be made. The small amounts of sample material necessary allow complete mtDNA sequences to be determined even from very valuable specimens. This approach makes it possible to sample the mitochondrial genomes of Pleistocene animals as widely as those of extant species, and provides the opportunity to answer detailed questions about the structure and history of extinct populations.

METHODS

Detailed descriptions of the methods used to date the mammoth bone and extract, amplify and sequence its DNA, together with the phylogenetic methods used, are provided in the Supplementary Information.

DNA extraction and amplification. For DNA extraction we used 747 mg bone powder from a mammoth bone found in Berelekh, Yakutia (Supplementary Information), yielding 70 µl of extract. We used 1.5 µl of the DNA extract for each multiplex PCR reaction, in a total volume of 20 µl. After 27 cycles of PCR, the appropriate multiplex product was diluted and 0.625% of this material was used for each of 46 individual amplifications. Amplification products of the correct size were cloned using the TOPO TA cloning kit (Invitrogen), and a minimum of three clones were sequenced on an ABI3730 capillary sequencer (Applied Biosystems).

We designed new primers for those fragments that gave only weak (or no) amplification products in the first attempts and for which the sequences of adjacent fragments showed differences from the elephant sequence in the primer sites. These new primers, together with the previously successful primers, were used to amplify the remaining segments of the mammoth mtDNA and to replicate all positions at least twice from independent primary amplifications (see also Supplementary Information). Extraction protocols and PCR conditions for the laboratories in Cambridge and London are available in the Supplementary Information.

Phylogenetic analyses. We initially aligned the mtDNA sequence of the mammoth to the corresponding sequences from Asian (*Elephas maximus*) and African (*Loxodonta africana*) elephant and dugong (*Dugong dugon*), excluding the control region in all analyses. Applying Modeltest²⁵ on this alignment, we obtained a general time-reversible substitution model with gamma-distributed substitution rates across sites. We estimated neighbour-joining and maximum-likelihood trees correcting for multiple substitutions using this substitution model and maximum parsimony trees using the program package PAUP²⁶, and bayesian trees using MrBayes²⁷. Depending on the tree-building method and outgroup, we recovered both the Asian and African elephant as sister taxa of the mammoth. We therefore added the mtDNA sequence from hyrax (*Procavia capensis*) as additional outgroup in an attempt to gain resolution. Again, we could not resolve the phylogeny (Supplementary Table S2).

Our analyses indicate that both dugong and hyrax are too distantly related to the mammoth and African and Asian elephants to resolve the phylogeny. However, by assuming a molecular clock, rooted phylogenies can be obtained without outgroups. We first tested the clock assumption by applying a simple test that uses only three taxa¹⁹. We also performed a clock test based on a likelihood ratio test with the program TREE-PUZZLE²⁸, using the Hasegawa–Kishino–Yano (HKY) model of DNA sequence evolution²⁹ and a gamma-distribution using eight gamma rate categories to model substitution rate heterogeneity

among sites. Neither of the two tests could reject the clock assumption. We subsequently used a parsimony method developed to estimate the phylogenetic relationship of three sequences that evolve under a molecular clock²⁰ and a likelihood ratio test (Supplementary Information) in order to resolve the relationship between mammoth, African and Asian elephant.

The parsimony method uses two statistics, *C* and *S* in Felsenstein's notation²⁰. Both are based on the number of positions at which one of the three sequences differs from the other two (which are identical to each other). *C* is the highest number of such positions obtained from one of the three possible comparisons n_1 , n_2 and n_3 (n_1 , species A differs from species B and C; n_2 , B differs from A and C; n_3 , C differs from A and B). *S* is described by $n_1 - n_2$, given that n_1 and n_2 are the largest and second largest numbers obtained for n , respectively. Excluding the control region, 403 (n_1) positions support a sister-group relationship between the mammoth and Asian elephant, 357 (n_2) positions support a sister-group relationship between the mammoth and the African elephant, and 339 (n_3) positions support a sister-group relationship between the two elephant species. This yields a *C* statistic of 403 ($C = n_1$) and an *S* statistic of 46 ($S = n_1 - n_2$)²⁰. On the basis of a Monte Carlo randomization test (10^6 simulations), the monophyly of the mammoth and the Asian elephant, to the exclusion of the African elephant, is significantly supported by both the *C* and the *S* statistic ($P = 0.032$ and $P = 0.035$, respectively).

To test whether the phylogenetic signals in the data are strong enough to warrant resolution of the sequence tree under a likelihood framework, we proceeded as follows. The likelihood of the data was assessed under two alternative models: (1) a simple model with only a single free parameter, corresponding to a star-like tree topology, and (2) a more complex model with two free parameters, resembling the resolved tree topology. A subsequent likelihood ratio test revealed that the more complex model, yielding the resolved sequence tree, explains the data significantly better, and thus rejects the simpler model of a star-like tree ($P < 0.01$).

Received 1 July; accepted 14 November 2005.

Published online 18 December 2005.

- Pääbo, S. *et al.* Genetic analyses from ancient DNA. *Annu. Rev. Genet.* **38**, 645–679 (2004).
- Lambert, D. M. *et al.* Rates of evolution in ancient DNA from Adelie penguins. *Science* **295**, 2270–2273 (2002).
- Cooper, A. *et al.* Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* **409**, 704–707 (2001).
- Haddrath, O. & Baker, A. J. Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis. *Proc. R. Soc. Lond. B* **268**, 939–945 (2001).
- Höss, M., Dilling, A., Currant, A. & Pääbo, S. Molecular phylogeny of the extinct ground sloth *Myodon darwini*. *Proc. Natl Acad. Sci. USA* **93**, 181–185 (1996).
- Loreille, O. *et al.* Ancient DNA analysis reveals divergence of the cave bear, *Ursus spelaeus*, and brown bear, *Ursus arctos*, lineages. *Curr. Biol.* **11**, 200–203 (2001).
- Yang, H., Golenberg, E. M. & Shoshani, J. Phylogenetic resolution within the Elephantidae using fossil DNA sequences from the American mastodon (*Mammot americanum*) as an outgroup. *Proc. Natl Acad. Sci. USA* **93**, 1190–1194 (1996).
- Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A. & Pääbo, S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Res.* **29**, 4793–4799 (2001).
- Bensasson, D., Zhang, D. X., Hartl, D. L. & Hewitt, G. M. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol.* **16**, 314–321 (2001).
- Greenwood, A. D. & Pääbo, S. Nuclear insertion sequences of mitochondrial DNA predominate in hair but not in blood of elephants. *Mol. Ecol.* **8**, 133–137 (1999).
- Greenwood, A., Capelli, C., Possnert, G. & Pääbo, S. Nuclear DNA sequences from late Pleistocene megafauna. *Mol. Biol. Evol.* **16**, 1466–1473 (1999).
- Noro, M., Masuda, R., Dubrovo, I. A., Yoshida, M. C. & Kato, M. Molecular phylogenetic inference of the woolly mammoth *Mammuthus primigenius*, based on complete sequences of mitochondrial cytochrome *b* and 12S ribosomal RNA genes. *J. Mol. Evol.* **46**, 314–326 (1998).
- Thomas, M. G., Hagelberg, E., Jone, H. B., Yang, Z. & Lister, A. M. Molecular and morphological evidence on the phylogeny of the Elephantidae. *Proc. R. Soc. Lond. B* **267**, 2493–2500 (2000).
- Debruyne, R., Barriel, V. & Tassy, P. Mitochondrial cytochrome *b* of the Lyakhov mammoth (Proboscidea, Mammalia): new data and phylogenetic analyses of Elephantidae. *Mol. Phylogenet. Evol.* **26**, 421–434 (2003).
- Ozawa, T., Hayashi, S. & Mikhelson, V. M. Phylogenetic position of mammoth and Steller's sea cow within Tethytheria demonstrated by mitochondrial DNA sequences. *J. Mol. Evol.* **44**, 406–413 (1997).
- Cummings, M. P., Otto, S. P. & Wakeley, J. Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* **12**, 814–822 (1995).

17. Shoshani, J. Understanding proboscidean evolution: a formidable task. *Trends Ecol. Evol.* **13**, 480–487 (1998).
18. Felsenstein, J. *Inferring Phylogenies* 196–221 (Sinauer Associates, Sunderland, 2004).
19. Tajima, F. Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135**, 599–607 (1993).
20. Felsenstein, J. Confidence-limits on phylogenies with a molecular clock. *Syst. Zool.* **34**, 152–161 (1985).
21. Tassy, P. in *Lothagam: The Dawn of Humanity in Eastern Africa* (eds Leakey, M. G. & Harris, J. M.) 331–358 (Columbia Univ. Press, New York, 2003).
22. Nei, M. in *Evolutionary Perspectives and the New Genetics* (eds Gershowitz, H., Rucknagel, D. L. & Tashian, R. E.) 133–147 (Alan R. Liss, New York, 1986).
23. Chen, F. C. & Li, W. H. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* **68**, 444–456 (2001).
24. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460 (1983).
25. Posada, D. & Crandall, K. A. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817–818 (1998).
26. Swofford, D. L. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)* (Sinauer Associates, Sunderland, 2003).
27. Ronquist, F. & Huelsenbeck, J. P. MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
28. Schmidt, H. A., Strimmer, K., Vingron, M. & Haeseler, A. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502–504 (2002).
29. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174 (1985).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the members of our laboratories for discussions and support, G. Khlopachev for providing the mammoth bone and K. Finstermeier for help with figure design. This work was supported by the Max Planck Society. J.L.P. and M.S. were supported by an NIH grant (to M.S.).

Author Information The complete mammoth mitochondrial DNA sequence has been deposited in GenBank under accession number DQ188829. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.H. (hofreiter@eva.mpg.de).

LETTERS

Localized maternal *orthodenticle* patterns anterior and posterior in the long germ wasp *Nasonia*

Jeremy A. Lynch¹, Ava E. Brent¹, David S. Leaf², Mary Anne Pultz² & Claude Desplan¹

The Bicoid (Bcd) gradient in *Drosophila* has long been a model for the action of a morphogen in establishing embryonic polarity¹. However, it is now clear that *bcd* is a unique feature of higher Diptera^{2,3}. An evolutionarily ancient gene, *orthodenticle* (*otd*), has a *bcd*-like role in the beetle *Tribolium*⁴. Unlike the Bcd gradient, which arises by diffusion of protein from an anteriorly localized messenger RNA^{1,5}, the *Tribolium* Otd gradient forms by translational repression of *otd* mRNA by a posteriorly localized factor. These differences in gradient formation are correlated with differences in modes of embryonic patterning. *Drosophila* uses long germ embryogenesis, where the embryo derives from the entire anterior–posterior axis, and all segments are patterned at the blastoderm stage, before gastrulation. In contrast, *Tribolium* undergoes short germ embryogenesis: the embryo arises from cells in the posterior of the egg, and only anterior segments are patterned at the blastoderm stage, with the remaining segments arising after gastrulation from a growth zone. Here we describe the role of *otd* in the long germband embryo of the wasp *Nasonia vitripennis*. We show that *Nasonia otd* maternal mRNA is localized at both poles of the embryo, and resulting protein gradients pattern both poles. Thus, localized *Nasonia otd* has two major roles that allow long germ development. It activates anterior targets at the anterior of the egg in a manner reminiscent of the Bcd gradient, and it is required for pre-gastrulation expression of posterior gap genes.

otd had been proposed as an ancestral anterior patterning gene in insects for two major reasons. First, it is highly conserved among animals and has an anterior patterning role in most phyla in which its function has been tested. Second, although it is quite distantly related to *bicoid*, Otd protein has a lysine at position 50 of its homeodomain that gives it the same DNA binding specificity as Bcd⁶.

Because a *Tribolium*-like system that forms an Otd gradient for patterning anterior structures based on a posteriorly localized source of translational repression would become increasingly inefficient as the germ rudiment extends more anteriorly, we sought to understand whether *otd* is a conserved anterior patterning factor in long germ insects that lack *bcd*.

To address this question, we examined the expression and function of *otd1* in the wasp *Nasonia*, which undergoes long germ development that seems to be morphologically similar to that of *Drosophila*, although early development takes longer^{7,8}. The *Nasonia otd1* gene is orthologous to *Tribolium otd1*, which has an early role in axis formation⁴. There is a second *otd* gene (*otd2*) in both species that is only expressed later in development⁹ (J.A.L. and C.D., manuscript in preparation).

We find that *Nasonia otd1* is expressed maternally in a surprising pattern. In early ovarian follicles, *otd1* mRNA is expressed in the nurse cells, and unexpectedly accumulates at the posterior of the oocyte (Fig. 1a). In later follicles, *otd1* mRNA remains localized at

the posterior of the oocyte but also begins accumulating at the anterior pole (Fig. 1b).

In pre-pole cell embryos, *otd1* mRNA is seen tightly localized at the anterior pole, whereas posteriorly localized mRNA becomes associated with the oosome, a structure that is thought to be an equivalent of germ plasm. As seen in Fig. 1c, this oosome-associated mRNA can migrate some distance from the posterior extremity, but returns to the pole just before the pole cells begin to form. After pole cell formation and nuclear migration to the surface of the embryo, *otd1* mRNA remains in a bi-polar expression pattern, but appears to

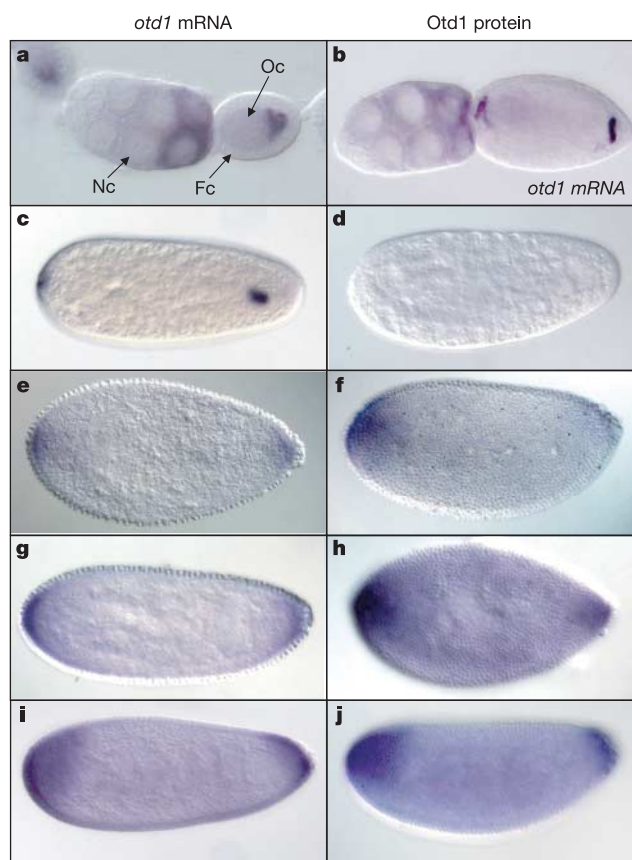


Figure 1 | *otd1* expression and localization. **a, b**, Ovarian *otd1* expression in early (**a**) and late (**b**) follicles. Nc, nurse cells; Fc, follicle cells; Oc, oocyte. **c, e, g, i**, Embryonic *otd1* mRNA expression in pre-pole cell (**c**), early (**e**) and late (**g**) syncytial blastoderm and cellular blastoderm (**i**) stages. **d, f, h, j**, Otd1 protein expression in pre-pole cell (**d**), early (**f**) and late (**h**) syncytial blastoderm, and cellular blastoderm (**j**) stages.

¹Center for Developmental Genetics, NYU Biology, 100 Washington Square East New York, New York 10003, USA. ²Western Washington University, Bellingham, Washington 98225, USA.

become less tightly localized to the cortex (Fig. 1c, e). Finally, after cellularization of the blastoderm, *otd1* mRNA is seen in cap-like domains at both the anterior and posterior poles (Fig. 1g).

The ovarian and early embryonic expression patterns of *Nasonia otd1* mRNA suggest that gradients of Otd1 protein may form by diffusion from localized sources of mRNA. An anterior Otd1 gradient would indicate possible convergent evolution of maternally localized mRNAs giving rise to gradients of K₅₀ homeoproteins in patterning the anterior of both *Drosophila* and *Nasonia*. A posterior Otd1 gradient would show that *Nasonia* has recruited *otd1* to perform a novel posterior patterning role not present in any of its known orthologues. To determine whether Otd1 gradients exist in *Nasonia*, we generated an antibody to *Nasonia* Otd1.

No Otd1 protein is seen before pole cell formation (Fig. 1d), indicating a mechanism of translational repression at this stage. *Nasonia* Otd1 protein is first detected when nuclei begin arriving at the surface of the embryo, just after pole cell formation. Notably, Otd1 is initially only seen at the anterior, where it forms an anterior to posterior gradient (Fig. 1f). A posterior to anterior gradient becomes visible at the posterior pole towards the end of the syncytial blastoderm stage (Fig. 1h). This pattern indicates a second level of translational repression specific to the posterior aspect of *Nasonia otd1* mRNA, and is consistent with the timing of posterior translational repression seen for *Nasonia hunchback* (*hb*)¹⁰. Interestingly, sequences that are similar to known Nanos response elements (NREs) are found in the 3' untranslated regions (UTRs) of both *Nasonia otd1* (gCGTTtcgccGcATTGTAcgag) and *hb*¹⁰ (where upper case letters indicate a match with the consensus sequence), indicating that posteriorly localized *nanos* may be responsible for preventing the translation of both mRNAs during early syncytial blastoderm stages. Finally, in the cellular blastoderm, Otd1 protein is seen in the anterior and posterior domain, mimicking the expression of mRNA.

The maternal localization of *otd1* mRNA along with subsequent Otd1 protein gradients are consistent with this gene acting as a morphogen at both poles of the *Nasonia* embryo. If this were indeed the case, when levels of Otd1 are reduced, the posterior borders of anterior Otd1 target genes should be shifted towards the anterior, whereas the anterior borders of posterior targets should be shifted posteriorly. To test the function of *otd1*, we adapted to *Nasonia* the parental RNA interference (pRNAi) technique first developed in *Tribolium*¹¹. We examined the expression of Otd1 in the offspring of *otd1* RNAi-injected mothers and observed varying amounts of reduction (see Supplementary Information), which correlates with variations in resulting phenotypes and effects on the expression patterns of the potential target genes *empty spiracles* (*ems*), *giant* (*gt*) and *hb*.

ems requires high levels of Bcd for expression in *Drosophila*¹². *Nasonia ems* is expressed in a very similar pattern to that of *Drosophila* (Fig. 2a). As a very anterior target of Otd1, it should be extremely sensitive to a reduction in Otd1 levels. Indeed, when *otd1* is knocked down, most embryos lose *ems* expression entirely, with only a few exhibiting reduced and anteriorly shifted expression (Fig. 2b, c).

Similar to *Drosophila*, the gap gene *gt* has two domains of expression in *Nasonia* (Fig. 2c). In most cases, when *otd1* is knocked down both *gt* stripes are shifted towards their respective poles (Fig. 2e); in the most strongly affected embryos, expression is lost at the anterior, whereas some residual expression is still seen at the posterior pole (Fig. 2f).

The gap gene *hb* responds to low levels of Bcd^{1,5,13} in flies and shows a broad anterior expression domain, as well as a posterior stripe¹⁴. *Nasonia hb* shows a similar pattern in the late blastoderm stages (Fig. 2g)¹⁰. In *otd1* RNAi embryos, the anterior *hb* domain shows a clear, although modest, anterior shift of its posterior boundary of expression (Fig. 2h, i). The degree to which *hb* is resilient to this

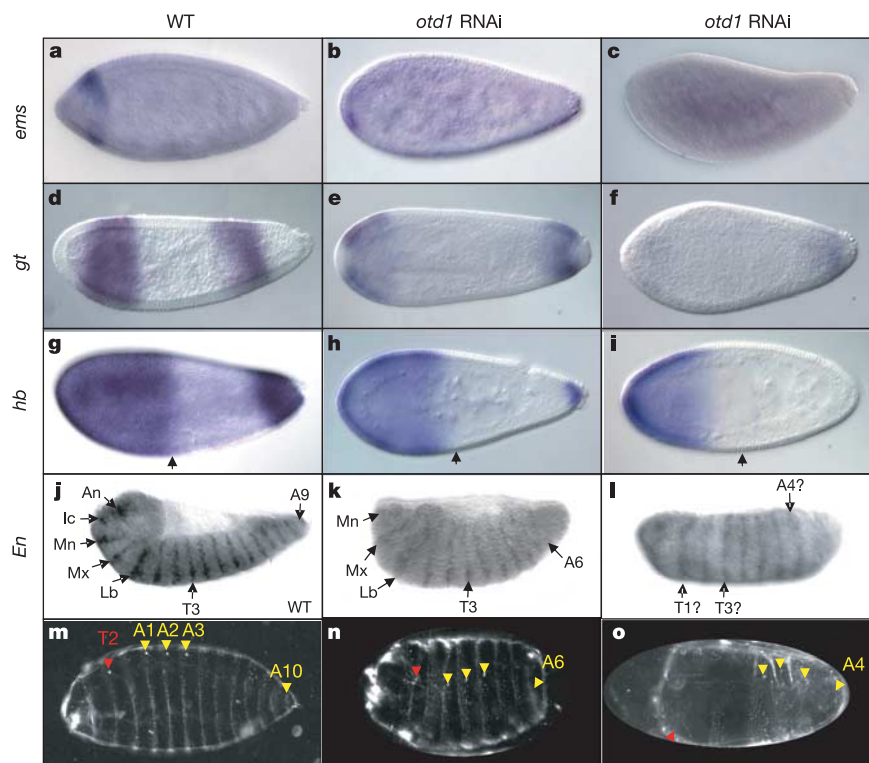


Figure 2 | Effects of *otd1* RNAi. **a–c**, *ems* expression in wild-type and *otd1* RNAi embryos. **d–f**, *gt* expression. **g–i**, *hb* expression (arrows indicate 50% egg length). **j–l**, Engrailed protein expression. The head segments have characteristic shapes. A, abdominal; An, antennal; Ic, intercalary; Lb, labial; Mn, mandibular; Mx, maxillary; T, thoracic. **m–o**, Cuticles of wild-type and

otd1 RNAi larvae. In wild-type larvae (**m**), mouthparts are visible anteriorly, and large spiracles are visible on the second thoracic (red arrow) and on the first three abdominal segments (yellow arrows). The left panel shows the wild-type pattern, whereas the middle panel shows moderate *otd1* RNAi phenotypes and the right panel severe *otd1* RNAi phenotypes.

knockdown is surprising, although it has been suggested that anterior zygotic *hb* can be activated by maternal *hb*¹⁰. Notably, although one aspect of zygotic *hb* expression is Bcd-dependent in *Drosophila*, this expression can be made dispensable as high levels of maternal *hb* can activate the zygotic function of *hb* required for thorax formation¹⁵.

In contrast to the modest effects on anterior *hb* expression, knocking down *otd1* expression has a marked effect on the posterior domain of *hb*. In *Drosophila*, this stripe is activated by *tailless*¹⁶, whereas in *Tribolium* it is not expressed until the latest stages of germband extension¹⁷. In *otd1* knocked down embryos, the anterior border of this stripe is either shifted to the extreme posterior pole of the embryo (Fig. 2h), or completely lost (Fig. 2i).

Knockdown of *otd1* results—presumably as a consequence of changes in expression of its targets—in the loss of both anterior and posterior segments: defects of varying severity are seen in both the expression of Engrailed protein (Fig. 2j–l and Table 1) and larval cuticle structures (Fig. 2m–o and Table 1). At the anterior end, Engrailed stripes and their corresponding cuticular segments are lost in an anterior to posterior progression, with RNAi phenotypes ranging from the loss of only the antennal stripe to the complete lack of head segments. At the posterior, Engrailed stripes and denticle belts appear to be lost in a posterior to anterior progression.

Because *hb* has been shown to cooperate with *bcd* in *Drosophila* and with *otd* in *Tribolium*^{4,18}, we performed double knockdown of *otd1* and *hb* (Table 1) to see whether a similar interaction exists in *Nasonia*. In the most severe cases, the entire anterior is lost, including several anterior abdominal segments. This is more severe than the combination of the most severe phenotypes seen by individual knockdown of either *otd1* or *hb*, or in the zygotic null *Nasonia hb* allele (*hb*^{headless})^{7,8,10} (Table 1; see also Supplementary Information), and the entire range of phenotypes is more severe. This indicates that *Nasonia otd1* and *hb* cooperate in anterior patterning. Although *hb* is also expressed and functions at the posterior, as in *Drosophila*, no increase in the severity or frequency of severe phenotypes is observed at the posterior in the double knockdown. The lack of synergy between *Otd1* and *Hb* at the posterior may allow different functions for the anterior and posterior *Otd1* gradients: *Otd1* acts with *Hb* at the anterior, whereas it might act alone or in combination with a

different factor (for example, Caudal; Fig. 3c) at the posterior to activate distinct sets of target genes.

These results show that *otd1* has broad roles in patterning both the anterior and posterior of the long germ *Nasonia* embryo. This

Table 1 | Effects of *otd1* RNAi on segmentation and synergy of *otd1* and *hb*

RNAi treatment	Remaining Engrailed head stripes (%)*						
	WT	Four	Three	Two	One	None	n
<i>otd1</i> RNAi	0	2	41	38	12	7	85
RNAi treatment	Cuticle phenotypes (%)*						
	WT	I	II	III	IV	V	n
<i>otd1</i> RNAi	0	91	7	2	0	0	120
<i>hb</i> RNAi	0	44	28	22	5	1	79
<i>hb otd1</i> RNAi	0	22	15	25	23	16	109
RNAi treatment	Remaining posterior abdominal segments (%)†						
	WT	≥Nine	Eight	Seven	Six	Five	n
<i>otd1</i> RNAi	0	11	27	33	22	7	107
<i>hb</i> RNAi	0	34	29	37	0	0	79
<i>hb otd1</i> RNAi	0	16	33	27	13	11	100

* Anterior defects caused by RNAi for *otd1*, *hb* and a combination of the two. Anterior *otd1* defects were quantified by examining the expression of Engrailed using the antibody 4D9 (ref. 26) to take advantage of the easily discernible shapes of the five head stripes (see also Fig. 2). The cuticle phenotypes are divided into five categories: I, weak head defects; II, no head structures; III, headless with thoracic defects; IV, no head or thorax; V, defects extending into anterior abdomen.
† Posterior RNAi phenotypes. There are ten abdominal denticle belts on the wild-type cuticle. The ≥nine abdominal segment class includes embryos with the normal number of denticle belts but defects in spacing.

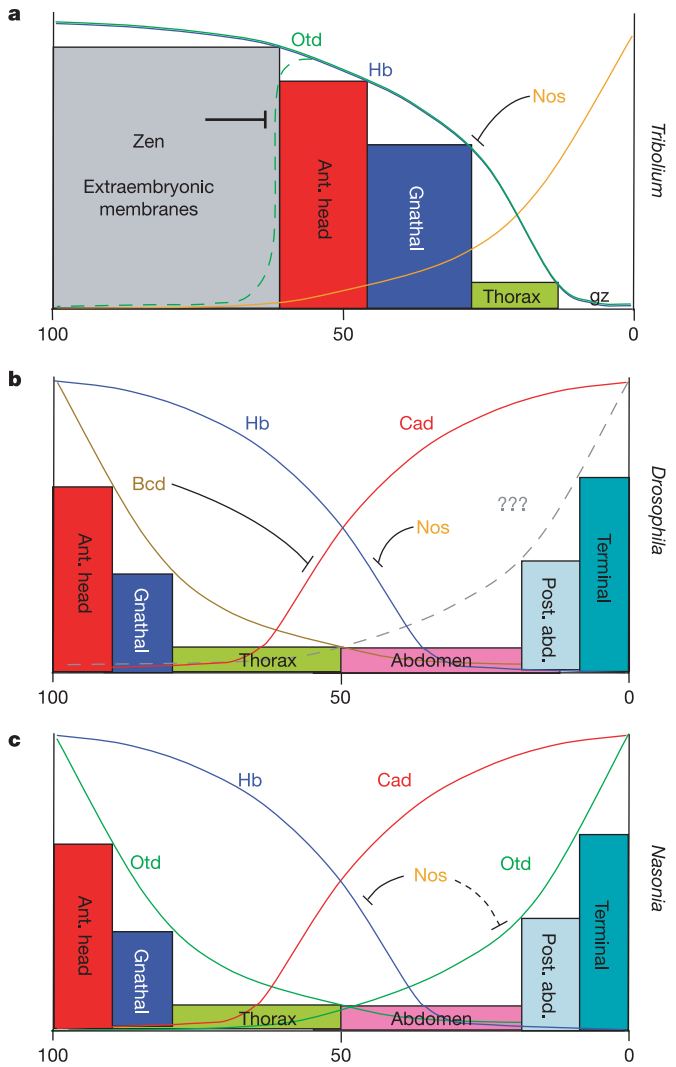


Figure 3 | Models for patterning the blastoderm of insect model systems. **a**, Schematic representation of patterning of the blastoderm stage of a short germ insect, based on what is known in *Tribolium*. In *Tribolium*, *Otd* and *Hb* gradients result from translational repression of ubiquitous mRNAs by a posteriorly localized factor, most likely *Nanos* (both of the genes have NREs in their 3' UTRs)^{4,17}. Genes responsible for patterning the head and thorax probably respond to these factors in a concentration-specific manner, much as they do to *Bcd* in *Drosophila*. The remainder of the embryo is patterned in the growth zone. *Tribolium* *Zerknullt* (*Zen*) represses embryonic fates in the anterior, and specifies the extraembryonic membranes²⁷. *gz*, growth zone. **b**, Summary of patterning in *Drosophila*. In this long germ embryo, anterior fates are established at the anterior pole as a result of a gradient of *Bcd* formed by diffusion of protein translated from a localized source of mRNA. It is not clear how the posterior genes are activated, but it may involve cooperation between *Bcd* and *Cad*, and possibly an additional posterior factor, such as *Tailless*. **c**, Schematic representation of the long germ blastoderm of *Nasonia*. The localization of *otd1* mRNA at the anterior pole of the egg probably results in a steep gradient of protein. This allows efficient patterning of the head and thorax in the anterior half of the egg, in synergy with *Hb*, which is restricted from the posterior pole, probably by *Nanos* (*Nanos* also seems to delay the formation of the posterior *Otd1* gradient, represented by the dashed line). *Nasonia otd1* is also localized to the posterior pole, which provides posterior positional information that is interpreted by downstream targets, perhaps by acting with *Cad*. *Nos*, *Nanos*.

provides insight into how a long germ type of patterning might have arisen from a short germ embryo. Major changes in patterning mechanisms at both the anterior and posterior ends of the embryo are required to pattern a long germ embryo. At the anterior, the egg fate map must be shifted so that the anterior (that is, head and thoracic) segments are patterned near the anterior pole. In addition, there must be a source of patterning information that is sufficiently strong to allow relatively tight spacing of these segments, in order to allow enough room at the posterior to pattern the abdomen. A system such as that found in *Tribolium* (Fig. 3a) does not seem compatible with this because the diffusion gradient arising from a posteriorly localized patterning centre will be shallowest at the anterior, where it would carry the least patterning information. *Nasonia* solves this by localizing *otd1* mRNA at the anterior pole, which results in a protein gradient with the most patterning information at the anterior of the embryo. This allows genes for which the orthologues are expressed more posteriorly in *Tribolium* to be activated close to the anterior pole in *Nasonia* (Fig. 3c). This is similar to the strategy used in *Drosophila* with localized *bcd* transcript (Fig. 3b).

Posterior patterning poses another problem for long germ embryos. In short germ insects, genes patterning posterior segments (such as the posterior domains of *gt* and *hb*) are expressed after gastrulation in a 'growth zone', whereas long germ embryos must express them at the blastoderm stage, before gastrulation. Again, *Nasonia* solves this problem by localizing *otd1* mRNA, this time at the posterior pole. The posterior gradient of Otd1 seems to be required for proper expression of posterior gap genes before gastrulation (Fig. 3c). This is a divergent strategy compared to that used in *Drosophila*, where neither *bcd* nor *otd* mRNA is localized at the posterior, and it is not clear which gene(s) provides morphogenetic information at the posterior. It may be that Bcd function extends posteriorly, because it acts with Caudal to activate posterior *knirps*¹⁹ and *hairy* stripe seven²⁰. Alternatively, or in addition, the terminal system might act through the induction of *tailless*, which has been shown to have some characteristics of a morphogen at the posterior¹⁶.

Although our results give insights into mechanisms of developmental evolutionary change, it will be of interest to understand whether what is seen in *Nasonia* is part of a larger pattern. For example, both *Nasonia* and *Drosophila* use the localization of mRNA encoding K₅₀ homeoprotein (Otd1 or Bcd) to pattern the anterior of the egg. Is this a general strategy for long germband patterning? On the other hand, there is no clear parallel between the posterior patterning systems of *Nasonia* and *Drosophila*. Is either strategy more common in other long germband embryos? To address this, the mechanisms used in other insects that have evolved long germband embryos must be elucidated. A number of key long germband taxa, including flies that lack *bcd*³, a mosquito²¹, moths²² and a beetle²³ have been used as laboratory organisms, and the application of further gene expression and functional analyses to these organisms should allow the placement of the *Nasonia* patterning system in a larger context.

METHODS

Nasonia orthologues of *otd1*, *ems* and *gt* were cloned using the following strategy: orthologues of these genes were identified in GenBank, and aligned using the MEGALIGN program in the DNASTAR package. Degenerate primers corresponding to conserved regions were designed using the CODEHOP algorithm²⁴. Sequences obtained by degenerate polymerase chain reaction (PCR) were extended in the 5' and 3' directions by RACE-PCR using the SMART RACE kit (Clontech).

For single RNAi, a 1 µg µl⁻¹ solution of double stranded (ds)RNA was injected into young female pupae, and the embryos laid by the injected mothers were examined. For double RNAi experiments, dsRNA for both genes were at final concentration of 0.75 µg µl⁻¹. Negative controls of distilled water and GFP dsRNA caused no defects.

Embryos for *in situ* hybridization and antibody experiments were fixed in

formaldehyde-saturated heptane for 30 min, affixed to double-sided tape and the vitelline membranes removed by hand under PBS.

Digoxigenin-labelled probes were generated using run-off transcription, using either T7 or SP6 RNA polymerase, of cloned fragments of the genes analysed in this work. *In situ* hybridization was done using standard protocols²⁵.

The affinity purified *Nasonia* Otd1 antibody was raised against the peptide CSPPRAKEAPGNP (Sigma-Genosys), used at a concentration of 1:100, and detected using an alkaline phosphatase conjugated anti-rabbit secondary antibody at 1:500 dilution. The monoclonal Engrailed antibody 4D9 was a gift from N. Patel, used at 1:10 dilution, and was detected using an alkaline phosphatase conjugated anti-mouse secondary antibody at 1:500 dilution. Cuticles were incubated at 65 °C overnight in 90% lactic acid/10% ethanol under a coverslip and photographed using dark-field microscopy.

Received 23 August; accepted 26 October 2005.

1. Driever, W. & Nusslein-Volhard, C. A gradient of *bicoid* protein in *Drosophila* embryos. *Cell* **54**, 83–93 (1988).
2. Dearden, P. & Akam, M. Developmental evolution: Axial patterning in insects. *Curr. Biol.* **9**, R591–R594 (1999).
3. Stauber, M., Prell, A. & Schmidt-Ott, U. A single Hox3 gene with composite *bicoid* and *zerknüllt* expression characteristics in non-Cyclorrhaphan flies. *Proc. Natl Acad. Sci. USA* **99**, 274–279 (2002).
4. Schroder, R. The genes *orthodenticle* and *hunchback* substitute for *bicoid* in the beetle *Tribolium*. *Nature* **422**, 621–625 (2003).
5. Struhl, G., Struhl, K. & Macdonald, P. M. The gradient morphogen *bicoid* is a concentration-dependent transcriptional activator. *Cell* **57**, 1259–1273 (1989).
6. Treisman, J., Gonczy, P., Vashishtha, M., Harris, E. & Desplan, C. A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell* **59**, 553–562 (1989).
7. Pultz, M. A., Pitt, J. N. & Alto, N. M. Extensive zygotic control of the anteroposterior axis in the wasp *Nasonia vitripennis*. *Development* **126**, 701–710 (1999).
8. Pultz, M. A. et al. A genetic screen for zygotic embryonic lethal mutations affecting cuticular morphology in the wasp *Nasonia vitripennis*. *Genetics* **154**, 1213–1229 (2000).
9. Li, Y. et al. Two *orthodenticle*-related genes in the short-germ beetle *Tribolium castaneum*. *Dev. Genes Evol.* **206**, 35–45 (1996).
10. Pultz, M. A. et al. A major role for zygotic *hunchback* in patterning the *Nasonia* embryo. *Development* **132**, 3705–3715 (2005).
11. Bucher, G., Scholten, J. & Klingler, M. Parental RNAi in *Tribolium* (Coleoptera). *Curr. Biol.* **12**, R85–R86 (2002).
12. Dalton, D., Chadwick, R. & McGinnis, W. Expression and embryonic function of *empty spiracles*: a *Drosophila* homeo box gene with two patterning functions on the anterior-posterior axis of the embryo. *Genes Dev.* **3**, 1940–1956 (1989).
13. Tautz, D. Regulation of the *Drosophila* segmentation gene *hunchback* by two maternal morphogenetic centres. *Nature* **332**, 281–284 (1988).
14. Bender, M., Horikami, S., Cribbs, D. & Kaufman, T. C. Identification and expression of the gap segmentation gene *hunchback* in *Drosophila melanogaster*. *Dev. Genet.* **9**, 715–732 (1988).
15. Wimmer, E. A., Carleton, A., Harjes, P., Turner, T. & Desplan, C. Bicoid-independent formation of thoracic segments in *Drosophila*. *Science* **287**, 2476–2479 (2000).
16. Margolis, J. S. et al. Posterior stripe expression of *hunchback* is driven from two promoters by a common enhancer element. *Development* **121**, 3067–3077 (1995).
17. Wolff, C., Sommer, R., Schroder, R., Glaser, G. & Tautz, D. Conserved and divergent expression aspects of the *Drosophila* segmentation gene *hunchback* in the short germ band embryo of the flour beetle *Tribolium*. *Development* **121**, 4227–4236 (1995).
18. Simpson-Brose, M., Treisman, J. & Desplan, C. Synergy between the *hunchback* and *bicoid* morphogens is required for anterior patterning in *Drosophila*. *Cell* **78**, 855–865 (1994).
19. Rivera-Pomar, R., Lu, X., Perrimon, N., Taubert, H. & Jackle, H. Activation of posterior gap gene expression in the *Drosophila* blastoderm. *Nature* **376**, 253–256 (1995).
20. La Rosee, A., Hader, T., Taubert, H., Rivera-Pomar, R. & Jackle, H. Mechanism and Bicoid-dependent control of *hairy* stripe 7 expression in the posterior region of the *Drosophila* embryo. *EMBO J.* **16**, 4403–4411 (1997).
21. Goltsev, Y., Hsiong, W., Lanzaro, G. & Levine, M. Different combinations of gap repressors for common stripes in *Anopheles* and *Drosophila* embryos. *Dev. Biol.* **275**, 435–446 (2004).
22. Kraft, R. & Jackle, H. *Drosophila* mode of metamerization in the embryogenesis of the lepidopteran insect *Manduca sexta*. *Proc. Natl Acad. Sci. USA* **91**, 6634–6638 (1994).
23. Patel, N. H., Condron, B. G. & Zinn, K. Pair-rule expression patterns of *even-skipped* are found in both short- and long-germ beetles. *Nature* **367**, 429–434 (1994).
24. Rose, T. M. et al. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences. *Nucleic Acids Res.* **26**, 1628–1635 (1998).

25. Tautz, D. & Pfeifle, C. A non-radioactive *in situ* hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene *hunchback*. *Chromosoma* **98**, 81–85 (1989).
26. Patel, N. H. *et al.* Expression of engrailed proteins in arthropods, annelids, and chordates. *Cell* **58**, 955–968 (1989).
27. van der Zee, M., Berns, N. & Roth, S. Distinct functions of the *Tribolium* *zerknüllt* genes in serosa specification and dorsal closure. *Curr. Biol.* **15**, 624–636 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank the Desplan and Small laboratories for support during this work, and J. Blau, E. Mazzoni, E. Olesnick, M. Siegal, S. Small, J. Treisman, D. Vasilaukas and E. Wimmer for critical reading of this manuscript. This work was supported by a NIH grant to C.D.; J.A.L. was

supported by a NIH Training Grant and a Dean Dissertation Fellowship from NYU. This investigation was conducted in a facility constructed with support from a Research Facilities Improvement grant from NCRR, NIH.

Author Contributions J.A.L. and C.D. conceived and designed the experiments. J.A.L. performed the experiments and generated data for the figures. A.E.B. cloned *Nasonia giant* and provided the probe. M.A.P. and D.S.L. cloned *Nasonia hunchback*, provided valuable reagents and were instrumental in initiating the project. J.A.L. and C.D. wrote the paper.

Author Information The sequences reported here are deposited in GenBank under the following accession numbers: *otd1*, AY684810; *ems*, AY684808; *giant*, DQ250085. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.D. (cd38@nyu.edu).

Gamma-band synchronization in visual cortex predicts speed of change detection

Thilo Womelsdorf^{1*}, Pascal Fries^{1,2*}, Partha P. Mitra³ & Robert Desimone^{4,5}

Our capacity to process and respond behaviourally to multiple incoming stimuli is very limited. To optimize the use of this limited capacity, attentional mechanisms give priority to behaviourally relevant stimuli at the expense of irrelevant distractors. In visual areas, attended stimuli induce enhanced responses and an improved synchronization of rhythmic neuronal activity in the gamma frequency band (40–70 Hz)^{1–11}. Both effects probably improve the neuronal signalling of attended stimuli within and among brain areas^{1,12–16}. Attention also results in improved behavioural performance and shortened reaction times. However, it is not known how reaction times are related to either response strength or gamma-band synchronization in visual areas. Here we show that behavioural response times to a stimulus change can be predicted specifically by the degree of gamma-band synchronization among those neurons in monkey visual area V4 that are activated by the behaviourally relevant stimulus. When there are two visual stimuli and monkeys have to detect a change in one stimulus while ignoring the other, their reactions are fastest when the relevant stimulus induces strong gamma-band synchronization before and after the change in stimulus. This enhanced gamma-band synchronization is also followed by shorter neuronal response latencies on the fast trials. Conversely, the monkeys' reactions are slowest when gamma-band synchronization is high in response to the irrelevant distractor. Thus, enhanced neuronal gamma-band synchronization and shortened neuronal response latencies to an attended stimulus seem to have direct effects on visually triggered behaviour, reflecting an early neuronal correlate of efficient visuo-motor integration.

Two monkeys were trained to perform a change detection task while spikes and local field potentials (LFPs) were recorded from several electrodes in area V4 (Fig. 1a, b; see Methods for details), an area that is strongly modulated by attention^{3–5}. Local synchronization was assessed by the coherence spectrum between spike trains and LFPs, as well as the power spectrum of the LFPs. We previously found that visual stimuli induced gamma-band synchronization in V4, which was enhanced when the stimulus was attended⁴ (Fig. 1c). Here we use data from our previous study in a new analysis that focuses on the behavioural reaction times to the stimulus change and on any associated changes in power, coherence and firing rates, time resolved in successive 10-ms steps with a sliding analysis window of ± 125 ms.

We first calculated LFP power ($n = 64$ recording sites) and spike–LFP coherence ($n = 244$ pairs of recordings sites) in the gamma frequency band (40–72 Hz), as well as firing rates ($n = 61$ recording sites) separately for the 25% trials with the slowest behavioural reactions and the 25% trials with the fastest reactions. Spike–field coherence from one pair of recording sites is shown in Fig. 2. Both in this example and across the set of recordings, we found that trials

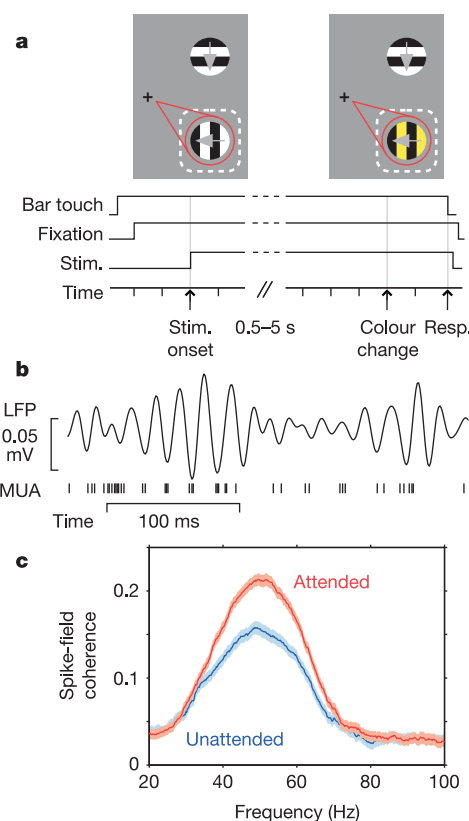


Figure 1 | Stimuli, behavioural model and examples of gamma-band synchronization and its modulation by attention. **a**, Monkeys started a trial by touching the bar and directing gaze to the fixation point. After a baseline period, two stimuli were presented, one inside the receptive field of the recorded neurons (broken rectangle) and one outside. On separate trials and before stimulus presentation, monkeys were cued to attend to one of the stimulus locations (red spotlight) and ignore the other. They were rewarded for releasing the bar on a subtle colour change in the cued stimulus while ignoring equally likely changes of the uncued stimulus. Changes could occur at any moment between 0.5 and 5 s after stimulus onset. **b**, Example of rhythmic multiunit activity (MUA) that is synchronized to gamma-band oscillations in the LFP. Multiunit activity and LFP were recorded from two separate electrodes. **c**, Example of spike–field coherence as a function of frequency for one pair of recording sites. The red (blue) lines show data recorded when the monkey directed attention into (away from) the receptive field of the recorded neurons. Shaded regions indicate ± 1 s.e.m.

¹FC Donders Centre for Cognitive Neuroimaging, Radboud University Nijmegen, 6525 EN Nijmegen, The Netherlands. ²Department of Biophysics, Radboud University Nijmegen, 6525 EZ Nijmegen, The Netherlands. ³Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. ⁴Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892, USA. ⁵McGovern Institute for Brain Research at MIT, Cambridge, Massachusetts 02139, USA.

*These authors contributed equally to this work.

leading to fast reaction times contained more gamma-band power (Fig. 3a) and gamma-band spike-field coherence (Fig. 3b) during epochs before and after the stimulus change event. The earliest significant change was found for spike-field coherence at 350 ms preceding the change event, whereas gamma-band power in the LFP was significantly enhanced ~ 125 ms before the stimulus change. Because of the ± 125 -ms width of the analysis window, we cannot precisely localize the beginning of these effects in time, although it is clear that they begin before the change event itself. Thus, enhanced gamma-band synchrony seems to lead to faster behavioural responses. Spike rates analysed with the same ± 125 -ms analysis window that was used for the spectral analysis showed a significant increase beginning 50 ms before the change event in trials leading to fast reaction times (Fig. 3c). Given the width of the analysis window, this could be an enhanced visual response to the change event itself.

In addition to these neuronal effects before the change event, all measures showed stronger gamma-band synchrony and spike rates in response to the change event on fast trials. On average, fast reactions were associated with a relative gamma-band power enhancement of 14.1% in the analysis windows centred between 0 and 75 ms after the change event, and an average spike rate enhancement of 8%. We did not consider effects beyond 75 ms after the change event, because the ± 125 -ms analysis window in this case would include the time of the earliest behavioural responses, which could be as fast as 200 ms (see Supplementary Fig. 1). We also computed spike rates on the basis of a narrower analysis window (gaussian with s.d. of 10 ms) and found that trials with fast reactions had shorter response onset latencies to the change event (~ 10 – 20 ms), and an enhanced evoked response to the stimulus in the period from 40 to 130 ms after the stimulus change (Fig. 3d).

The response and coherence differences found on fast versus slow trials could be selective to the attended stimulus, or they could be due to a general increase in arousal, or alertness, on trials with fast reaction times¹⁷. If the latter is true, we would expect enhanced gamma-band power and coherence also on trials with rapid behavioural responses to the stimulus outside the receptive field of the recorded neurons. We therefore computed power, coherence and firing rates in response to the stimulus inside the receptive field when the monkey was attending to a stimulus outside the receptive field and around the time of the change of this attended stimulus outside the receptive field. In contrast to the results when the monkey responded to the stimulus change inside the receptive field, gamma-band power and coherence at the unattended location were significantly reduced throughout most of the pre-change period

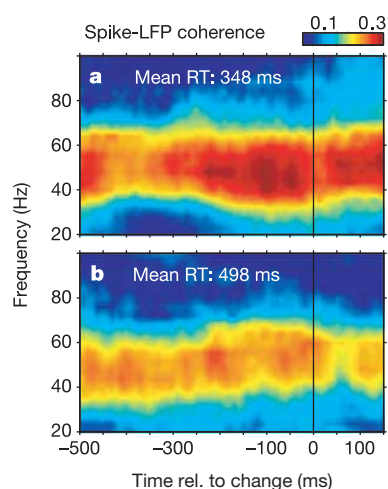


Figure 2 | Spike-field coherence from one pair of recording sites. Shown are averages over the 25% of trials with the fastest (a) and the slowest (b) reaction times.

when the monkeys responded quickly to the stimulus outside the receptive field (Fig. 3e, f). Spike rates were only marginally reduced for fast versus slow trials (Fig. 3g). Thus, the effects of attention on LFP power and coherence inside the receptive field were approximately reversed for fast versus slow responses to attended stimuli outside the receptive field (Fig. 3h), which is inconsistent with a general change in arousal.

These results suggest that when the monkey attends to one stimulus and ignores a distractor, the relative gamma-band coherence differences in response to the attended and ignored stimulus will, on average, predict fast versus slow reaction times well before the change event occurs. When the change event occurs, neuronal responses have a shorter latency and stronger gamma-band coherence on the faster trials. Some of the observed effects became apparent already in the first analysis window, 500 ms before the change event—that is, the time of stimulus onset for the shortest

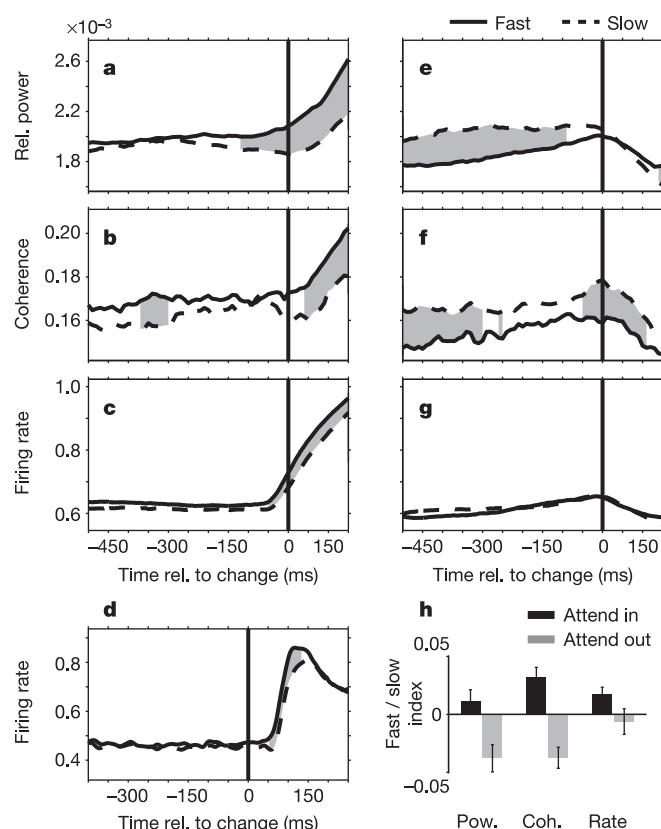


Figure 3 | Neuronal activity parameters in trials with fast and slow change detection. a–d, Time course of neuronal activity parameters induced by the attended stimulus inside the receptive fields around the change of the attended stimulus. a, Relative LFP power in the gamma band (40–72 Hz). b, Spike-field coherence in the gamma band. c, d, Firing rate. In a–c, analysis windows of ± 125 ms were used; in d, a gaussian kernel of 10-ms s.d. was used. Shown are grand averages calculated separately for the 25% of trials with the fastest (unbroken lines) and slowest (broken lines) behavioural responses. Grey shading indicates significance (two-sided paired *t*-test, $P < 0.05$ after multiple comparison correction). e–g, As a–c, but showing the time course of neuronal activity parameters induced by the ignored distractor, around the change of the attended stimulus that was outside the receptive fields. h, Comparison of neuronal activity parameters for fast versus slow trials (average over -500 to -125 ms preceding the change) when behavioural reports were in response to changes inside (black bars) or outside (grey bars) the receptive fields. Shown is the modulation index defined as $(P_{\text{fast}} - P_{\text{slow}})/(P_{\text{fast}} + P_{\text{slow}})$, where P_{fast} is one of the parameters investigated for trials with fast responses and P_{slow} is one for trials with slow responses. Error bars denote ± 1 s.e.m. Pow., power; Coh., coherence.

trials. To determine the earliest neural effects on reaction time, we therefore repeated the above analyses for a 250-ms window before stimulus onset, but we found no significant effects in that interval.

To test directly whether trial-by-trial fluctuations of synchrony could predict the speed of change detection on a single-trial basis, we performed a correlation analysis. For this, we correlated trial-by-trial variations in reaction time with trial-by-trial variations in coherence or power and firing rate (see Methods). To obtain an overview of the spectral specificity of the neuronal response changes, we extended the time-resolved analysis to frequencies of 8–100 Hz. Consistent with the preceding analysis, we found that short reaction times were predicted by enhanced power and coherence in the gamma-frequency band (40–72 Hz) in time epochs preceding the stimulus change by several hundred milliseconds (Fig. 4a, b). Also shortly before and after the change event, behavioural response times were predicted by the degree of synchrony and neuronal response magnitude. Correlations of firing rate (analysed within the same sliding ± 125 -ms analysis window used above) and reaction time were significant from 40 ms before the change and onwards (Fig. 4c). In contrast to firing rate and gamma-band modulation, we observed reduced power in the alpha (10 Hz) and beta (15 Hz) frequency bands for fast trials in a restricted time window starting 80 ms before the stimulus change (Fig. 4a).

Across measures, the strongest correlations with reaction times were evident at a time immediately after the stimulus change. To analyse this effect for single recording sites, we calculated Z-scores of the correlation of reaction time with gamma-band power and spike

rate averaged over the analysis windows centred between 0 and 75 ms after the change event. The distribution of Z-scores for LFP power and firing rate is strongly biased towards negative values (see Supplementary Fig. 2 and Supplementary Table). Moreover, the negative reaction time correlations are clearly evident in both monkeys, which is noteworthy because the two monkeys showed a different trend in reaction time speed as a function of time in trial (namely, a slightly increasing/decreasing response speed with time in trial; see Supplementary Fig. 1 and Supplementary Note). Thus, while the two monkeys showed weak but opposite trends in reaction times across the trial, they both showed the same patterns of correlation between reaction times and spectral power, firing rate and coherence.

Our results show that the degree of neuronal synchrony in the gamma-frequency band in time intervals preceding and following a behaviourally relevant sensory change can predict the speed of behavioural responses to that change. These effects are reversed for responses to stimulus changes outside the receptive field.

The neuronal processes in visual area V4, studied here, constitute only one link in the processing chain from the stimulus change event to its behavioural report. Oscillatory synchronization within and between other structures along this way has been described and probably also has a functional role^{16,18–24}. Enhanced gamma-band synchronization in V4 could be directly involved in the detection process and/or in the signalling of the detection achieved in V4 or a preceding area. In both cases, enhanced gamma-band synchronization among the recorded neurons probably subserves rapid and reliable signalling mechanistically, because it results in efficient summation of postsynaptic potentials^{13,25,26}.

A reported effect of gamma-band synchronization during the onset of visual stimuli is enhanced synchronization and shortened response latencies of the first spikes on stimulus presentation²⁷. If this finding is extended to stimulus changes, in addition to stimulus onsets, it suggests that there is a mechanistic link between enhanced gamma-band synchrony around the time of the sensory change and the more rapid, shifted response latency that we observe. Taken together, our results suggest that the enhanced precise synchronization is instrumental in subserving a rapid and reliable transmission of information about sensory changes to the postsynaptic targets and thus ultimately triggers enhanced detection efficiency.

METHODS

Procedures were done in accordance with NIH guidelines and were approved by the National Institute of Mental Health (NIMH) Intramural Animal Care and Use Committee. Simultaneous recordings of spikes and local field potentials were made from four to eight electrodes in visual areas V4 in two hemispheres of two monkeys. In total, multiunit recordings were obtained from 61 sites (39 and 22 in monkey A and B, respectively) and LFP recordings from 64 sites (40 and 24, respectively). Power and coherence spectra were assessed for windows of ± 125 ms, moved over the data in steps of 10 ms from 500 ms before to 150 ms after the stimulus change. Neuronal activity parameters were compared between trials with the 25% fastest and trials with the 25% slowest response times of individual recording sessions. The mean reaction time for the fast and slow trials was 346 ms and 490 ms, respectively (median, 359 and 485 ms, respectively). To determine the predictive capability of neuronal activity parameters for reaction time, we calculated the Pearson correlation coefficient between the trial-by-trial variations in reaction times and the trial-by-trial variations in power, coherence and spike rate. See Supplementary Methods for more information.

Received 1 June; accepted 23 September 2005.

Published online 21 December 2005.

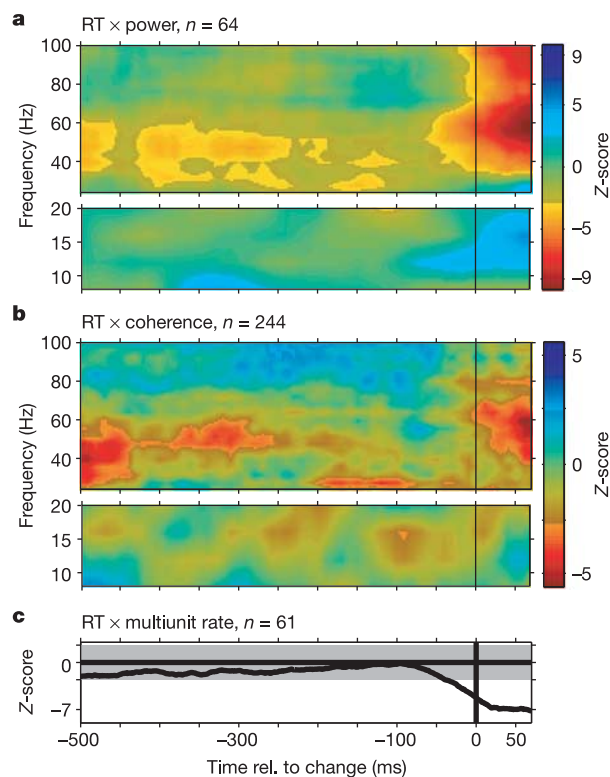


Figure 4 | Trial-by-trial prediction of reaction times by neuronal activity. Shown are Z-scores of correlation coefficients pooled across monkeys and recordings sites (a, c) or pairs of recording sites (b) on the basis of analysis windows of ± 125 ms. Note inverted colour scale. Grey shading highlights significance (99% confidence). Results are shown separately for lower and higher frequencies because different spectral concentrations were used (see Methods). a, LFP power versus reaction times. b, Spike-field coherence versus reaction times. c, Firing rates versus reaction times.

- Engel, A. K., Fries, P. & Singer, W. Dynamic predictions: oscillations and synchrony in top-down processing. *Nature Rev. Neurosci.* **2**, 704–716 (2001).
- Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).
- Motter, B. C. Neural correlates of attentive selection for colour or luminance in extrastriate area V4. *J. Neurosci.* **14**, 2178–2189 (1994).
- Fries, P., Reynolds, J. H., Rorie, A. E. & Desimone, R. Modulation of oscillatory neuronal synchronization by selective visual attention. *Science* **291**, 1560–1563 (2001).

5. Bichot, N. P., Rossi, A. F. & Desimone, R. Parallel and serial neural mechanisms for visual search in macaque area V4. *Science* **308**, 529–534 (2005).
6. Treue, S. & Maunsell, J. H. Attentional modulation of visual motion processing in cortical areas MT and MST. *Nature* **382**, 539–541 (1996).
7. Cook, E. P. & Maunsell, J. H. Dynamics of neuronal responses in macaque MT and VIP during motion detection. *Nature Neurosci.* **5**, 985–994 (2002).
8. Steinmetz, P. N. *et al.* Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature* **404**, 187–190 (2000).
9. Taylor, K., Mandon, S., Freiwald, W. A. & Kreiter, A. K. Coherent oscillatory activity in monkey area V4 predicts successful allocation of attention. *Cereb. Cortex* **15**, 1424–1437 (2005).
10. Gruber, T., Müller, M. M., Keil, A. & Elbert, T. Selective visual–spatial attention alters induced gamma band responses in the human EEG. *Clin. Neurophysiol.* **110**, 2074–2085 (1999).
11. Martinez, A. *et al.* Involvement of striate and extrastriate visual cortical areas in spatial attention. *Nature Neurosci.* **2**, 364–369 (1999).
12. Salinas, E. & Sejnowski, T. J. Impact of correlated synaptic input on output firing rate and variability in simple neuronal models. *J. Neurosci.* **20**, 6193–6209 (2000).
13. Salinas, E. & Sejnowski, T. J. Correlated neuronal activity and the flow of neural information. *Nature Rev. Neurosci.* **2**, 539–550 (2001).
14. Tiesinga, P. H. & Sejnowski, T. J. Rapid temporal modulation of synchrony by competition in cortical interneuron networks. *Neural Comput.* **16**, 251–275 (2004).
15. Börgers, C., Epstein, S. & Kopell, N. J. Background gamma rhythmicity and attention in cortical local circuits: A computational study. *Proc. Natl Acad. Sci. USA* **102**, 7002–7007 (2005).
16. Schoffelen, J. M., Oostenveld, R. & Fries, P. Neuronal coherence as a mechanism of effective corticospinal interaction. *Science* **308**, 111–113 (2005).
17. Munk, M. H., Roelfsema, P. R., König, P., Engel, A. K. & Singer, W. Role of reticular activation in the modulation of intracortical synchronization. *Science* **272**, 271–274 (1996).
18. Lee, D. Coherent oscillations in neuronal activity of the supplementary motor area during a visuomotor task. *J. Neurosci.* **23**, 6798–6809 (2003).
19. Riehle, A., Grün, S., Diesmann, M. & Aertsen, A. Spike synchronization and rate modulation differentially involved in motor cortical function. *Science* **278**, 1950–1953 (1997).
20. Mehring, C. *et al.* Inference of hand movements from local field potentials in monkey motor cortex. *Nature Neurosci.* **6**, 1253–1254 (2003).
21. Pesaran, B., Pezaris, J. S., Sahani, M., Mitra, P. P. & Andersen, R. A. Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nature Neurosci.* **5**, 805–811 (2002).
22. Scherberger, H., Jarvis, M. R. & Andersen, R. A. Cortical local field potential encodes movement intentions in the posterior parietal cortex. *Neuron* **46**, 347–354 (2005).
23. Gonzalez Andino, S. L., Michel, C. M., Thut, G., Landis, T. & Grave de Peralta, R. Prediction of response speed by anticipatory high-frequency (gamma band) oscillations in the human brain. *Hum. Brain Mapp.* **24**, 50–58 (2005).
24. Gray, C. M., König, P., Engel, A. K. & Singer, W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**, 334–337 (1989).
25. Azouz, R. & Gray, C. M. Adaptive coincidence detection and dynamic gain control in visual cortical neurons in vivo. *Neuron* **37**, 513–523 (2003).
26. MacLeod, K., Bäcker, A. & Laurent, G. Who reads temporal information contained across synchronized and oscillatory spike trains? *Nature* **395**, 693–698 (1998).
27. Fries, P., Neuenschwander, S., Engel, A. K., Goebel, R. & Singer, W. Rapid feature selective neuronal synchronization through correlated latency shifting. *Nature Neurosci.* **4**, 194–200 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. H. Reynolds, A. E. Rorie and A. F. Rossi for help with experiments; and R. Oostenveld and M. Bauer for discussions. This research was supported by grants from the Human Frontier Science Program Organization (to P.F. and P.P.M.), The Netherlands Organization for Scientific Research and The Volkswagen Foundation (to P.F.), the NIH (to P.P.M.) and the Intramural Program of the NIH, NIMH.

Author Contributions T.W. and P.F. contributed equally to this work. P.F. conceived and performed the experiment. T.W., P.F. and P.P.M. conceived and performed the data analysis. P.F., T.W. and R.D. wrote the paper.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to T.W. (t.womelsdorf@fcdonders.ru.nl).

Whole-cell patch-clamp measurements of spermatozoa reveal an alkaline-activated Ca^{2+} channel

Yuriy Kirichok¹, Betsy Navarro¹ & David E. Clapham¹

In mammals, sperm cells become motile during ejaculation and swim up the female reproductive tract. Before fertilization and to overcome various barriers, their motility must be hyperactivated, a motion that is characterized by vigorous asymmetric tail beating¹. Hyperactivation requires an increase in calcium in the flagella, a process that probably involves plasmalemmal ion channels^{2–8}. Numerous attempts in the past two decades to understand sperm cell channels have been frustrated by the difficulty of measuring spermatozoan transmembrane ion currents^{2,3,9–16}. Here, by using a simple approach to patch-clamp spermatozoa and to characterize whole-spermatozoan currents, we describe a constitutively active flagellar calcium channel that is strongly potentiated by intracellular alkalinization. This current is not present in spermatozoa lacking the sperm-specific putative ion channel protein, CatSper1. This plasma membrane protein of the six transmembrane-spanning ion channel superfamily is specifically localized to the principal piece of the sperm tail and is required for sperm cell hyperactivation and male fertility^{4,5}. Our results identify CatSper1 as a component of the key flagellar calcium channel, and suggest that intracellular alkalinization potentiates CatSper current to increase intraflagellar calcium and induce sperm hyperactivation.

CatSper1, and three homologous proteins (CatSper2–CatSper4), comprise a family of sperm-specific proteins^{4,7,17}. In analogy to other members of the six transmembrane (6TM) channel family, CatSper proteins seem to be subunits of tetrameric voltage-gated cation channels. Attempts to express CatSper1 functionally in heterologous expression systems, with or without the other CatSper subunits, have been unsuccessful, perhaps owing to a requirement for the intraflagellar transport system of sperm. Here we have directly recorded the activity of ion channels from the plasma membrane of intact mouse sperm cells in whole-cell and perforated-patch configurations. Using these techniques, we have recorded a Ca^{2+} current (that we call CatSper or I_{CatSper}) that originates specifically from the principal piece of the sperm cell flagellum and is absent in *Catsper*^{−/−} mice.

After numerous unsuccessful attempts to form a giga-ohm seal between the patch pipette and the plasma membrane of the sperm cell head, midpiece and tail, we surmised that the tight link between the plasma membrane and stiff intracellular structures prevented seal formation. We circumvented this problem by applying the patch pipette to the small cytoplasmic droplet of the sperm (Fig. 1a and Supplementary Movie), a remnant of the precursor germ cell cytoplasm located on the midpiece of a mature sperm cell that is shed around the time of ejaculation¹⁸. We measured from sperm in which the cytoplasmic droplet was adjacent to the principal piece, in order to record from the most mature sperm in the corpus epididymis¹⁸. We cannot rule out the possibility that the sperm acquire

additional properties with further transport through the male reproductive tract. The plasma membrane of the cytoplasmic droplet is loosely attached to the stiff structural elements of the sperm cell flagellum, allowing formation of a tight seal with the pipette and subsequent break-in to generate a stable whole-cell configuration. In this configuration, the pipette had electrical and diffusion access to all compartments of the sperm cell (head, middle and principal

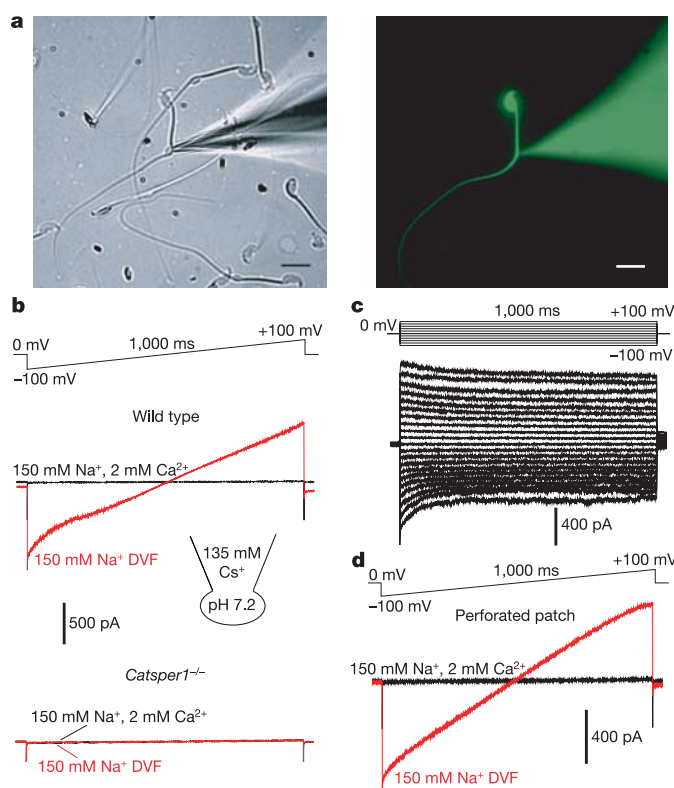


Figure 1 | Whole-cell patch-clamp recordings from intact spermatozoa. **a**, Left, the tight seal formed with the cytoplasmic droplet of the spermatozoon. Right, fluorescent image of the same spermatozoon after formation of the whole-cell configuration. Lucifer yellow (1 mM) diffused from the pipette throughout the interior of the spermatozoon in <5 s. Scale bar, 10 μm. **b**, Whole-cell currents elicited by a ramp from a holding potential (V_H) of 0 mV in DVF conditions from wild-type (top) and *Catsper1*^{−/−} (bottom) sperm. **c**, Representative I_{CatSper} in response to a step protocol from $V_H = 0$ mV in DVF conditions. **d**, I_{CatSper} recorded in perforated-patch mode of whole-cell recording.

¹Howard Hughes Medical Institute, Department of Cardiology, Children's Hospital and Department of Neurobiology, Harvard Medical School Enders 1309, Children's Hospital Boston 320 Longwood Avenue, Boston, Massachusetts 02115, USA.

pieces; Fig. 1a). Mean capacitance of the whole sperm cell membrane was 2.5 ± 0.2 pF (consistent with $\sim 250 \mu\text{m}^2$ of sperm membrane; $n = 346$).

Because CatSper1 has been presumed to be a cationic channel⁴, the bath and pipette solutions contained minimal concentrations of membrane-permeable anions. Generally, K^+ ions were also omitted to prevent K^+ currents. Under such conditions and in the presence of 2 mM Ca^{2+} in the bath solution, the sperm cell membrane showed a very low conductance (Fig. 1b). On switching to a divalent-free (DVF) solution, however, we recorded a robust, quasilinear monovalent current, consistent with results from fluorescent dye measurements^{19,20}. The current was carried by Na^+ in the inward direction and Cs^+ in the outward direction (Fig. 1b, c). This current, which we call I_{CatSper} , was absent in *Catsper1*^{-/-} sperm cells (Fig. 1b, $n = 25$). We tested whether whole-cell mode perfusion of the cell interior with the pipette solution washed out key intracellular components that kept the channel in its closed state. In amphotericin perforated-patch recording, a method that prevents washout²¹, I_{CatSper} was unaltered as compared with the normal whole-cell configuration (Fig. 1d). Our results indicate that the CatSper channels are constitutively active and that CatSper1 is required for this current. Preliminary evidence suggests that the CatSper channel comprises the CatSper1 protein and other CatSper family subunits⁶. Complete identification of the channel constituents will require further investigation.

Another property of the CatSper1 protein is its specific localization to the principal piece of the sperm flagellum⁴, raising the issue of whether I_{CatSper} can be detected only in the principal piece. We recorded ionic currents from sperm cell fragments containing only the principal and midpiece, or only the head and midpiece (Fig. 2). I_{CatSper} was recorded from the fragment of the sperm cell containing the principal piece and midpiece (Fig. 2a), but not from the fraction consisting of only the head and midpiece (Fig. 2b). Therefore, I_{CatSper} originates from the principal piece of the sperm flagellum. The large amplitude of these currents (~ 500 pA/pF) suggests that CatSper channels densely localize to the flagellar plasma membrane.

The monovalent current through CatSper channels was potentially decreased by addition of Ca^{2+} to the extracellular side of the membrane (Fig. 3a), indicating a markedly high-affinity binding site for Ca^{2+} in the pore region of CatSper channels, which is a definitive property of Ca^{2+} -selective channels^{22,23}. We found that monovalent I_{CatSper} was strongly potentiated by intracellular alkalinization (Supplementary Fig. 1), and thus reasoned that we might detect the smaller, but physiologically relevant Ca^{2+} current by increasing the pH of the pipette solution. As bath $[\text{Ca}^{2+}]_o$ (that is, $[\text{Ca}^{2+}]_o$) was gradually raised from nominally 0 to 10 mM at $\text{pH}_{\text{pip}} 8.0$, inward Ca^{2+} current increased (Fig. 3b). A similar inward Ba^{2+} current

could also be elicited by a gradual increase in $[\text{Ba}^{2+}]_o$ from 0 to 50 mM (Fig. 3b). Inward current was not observed when the principal cation was Mg^{2+} (up to 50 mM $[\text{Mg}^{2+}]_o$; Fig. 3b). In recordings from *Catsper1*^{-/-} sperm cells, neither Ca^{2+} ($n = 12$) nor Ba^{2+} ($n = 14$) currents were detected under these conditions. We conclude that I_{CatSper} is a Ca^{2+} -selective current under physiological conditions. This CatSper Ca^{2+} current gradually inactivated at $[\text{Ca}^{2+}]_o > 2$ mM, but fully recovered after incubation in a nominal Ca^{2+} -free bath solution for several minutes (data not shown). To circumvent this rundown, $[\text{Ba}^{2+}]_o$ was used in subsequent experiments.

The extremely small free cytoplasmic volume of the flagellum and high density of active CatSper channels in the flagellar plasma membrane lead to significant shifts in intraflagellar concentrations of ions. This is seen as outward Ba^{2+} current in Fig. 3b, resulting from the accumulation of intraflagellar Ba^{2+} . Like inward Ba^{2+} current, outward Ba^{2+} current was absent in *Catsper1*^{-/-} mice. The channel activity is highest at depolarized potentials, as seen in Supplementary Fig. 2 where a strongly outwardly rectifying I_{CatSper} was recorded in symmetrical 50 mM Ba^{2+} . This current was not detected in *Catsper1*^{-/-} sperm cells. Under physiological conditions, I_{CatSper} should be active, voltage sensitive and highly Ca^{2+} selective.

I_{CatSper} shows weak voltage dependence, but the range of its voltage sensitivity is shifted significantly by intracellular alkalinization. Voltage-dependent activation of I_{CatSper} was measured from tail currents elicited after repolarization from various test potentials (Fig. 4a). The resulting activation curve shows that little current is active at pH 6.0 in the physiological range of potentials (about -70 to 0 mV). On alkalinization to pH 7.5, however, the activation curve is shifted to more hyperpolarized potentials. The potential at which half the channels are activated ($V_{1/2}$) shifts (-76 mV, from +87 mV at pH 6.0 to +11 mV at pH 7.5 (Fig. 4b)). The slope factor (k) of ~ 30 indicates weak voltage sensitivity as compared with slope factors of ~ 4 for strongly voltage-sensitive ion channels.

Indirect methods have suggested that Ca^{2+} fluxes are regulated by intracellular pH in sperm and spermatogenic cells^{24,25}. The intracellular pH of resting spermatozoa is acidic ($\text{pH} \approx 6.5$)^{26,27}. We recorded very small I_{CatSper} Ba^{2+} currents when the pipette pH was

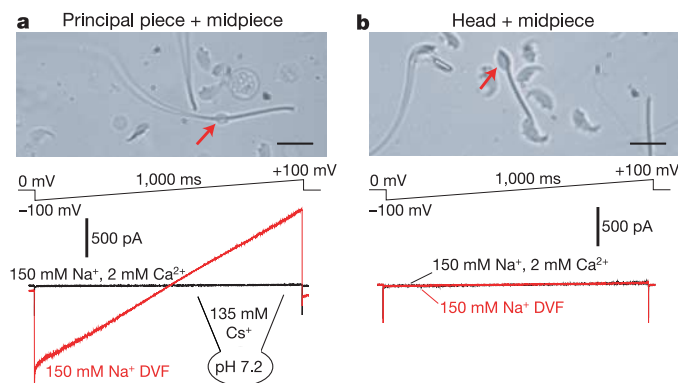


Figure 2 | I_{CatSper} originates from the principal piece of the sperm cell flagellum. **a**, Whole-cell recordings from a beheaded sperm cell containing only the mid- and principal piece (value $\pm$ s.e.m.: $C = 1.9 \pm 0.2$ pF; $n = 7$). **b**, Whole-cell recording from sperm cell head and midpiece ($C = 1.2 \pm 0.2$ pF; $n = 5$). Note the cytoplasmic droplet (arrow).

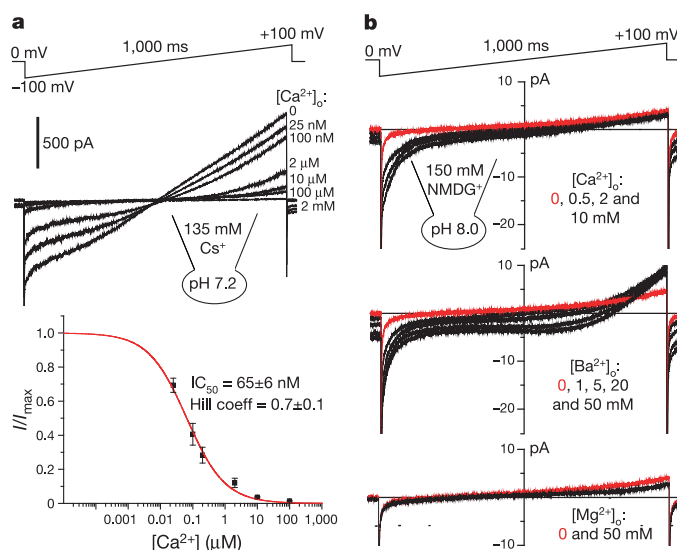


Figure 3 | I_{CatSper} is Ca^{2+} selective. **a**, Monovalent I_{CatSper} was potentially reduced in increasing $[\text{Ca}^{2+}]_o$. Top, representative I_{CatSper} traces elicited by a voltage ramp in the presence of different $[\text{Ca}^{2+}]_o$ ($V_H = 0$ mV). Bottom, relative I_{CatSper} amplitudes at -100 mV plotted against varying $[\text{Ca}^{2+}]_o$ (data are mean $\pm$ s.d.; $n = 8$). **b**, Divalent current elicited by a voltage ramp ($V_H = 0$ mV) with Ca^{2+} or Ba^{2+} as divalent ions. No inward current was detected in Mg^{2+} . '0' divalent concentration refers to nominal concentration.

6.5. As we increased the pH in the pipette, however, I_{CatSper} increased significantly (Fig. 4c). To buffer pH more accurately in the sperm cell, we replaced EGTA (which releases H^+ on binding Ba^{2+}) with BAPTA. Under these conditions and with reduced Ba^{2+} ($[\text{Ba}^{2+}]_o = 5 \text{ mM}$), I_{CatSper} increased from 2 ± 0.3 at $\text{pH}_{\text{pip}} 6.0$ to $12 \pm 1 \text{ pA}$ at $\text{pH}_{\text{pip}} 7.5$ ($n = 12$). I_{CatSper} could also be increased by intracellular alkalinization subsequent to bath application of $10 \text{ mM NH}_4\text{Cl}$ (Fig. 4d). Thus, capacitating conditions, which induce intracellular alkalinization in spermatozoa, should strongly potentiate I_{CatSper} . Notably, the amino terminus of CatSper1 protein is highly histidine rich (83 histidines in the 446-residue N terminus), suggesting that the N terminus has a role in sensing intracellular pH (ref. 4).

Cyclic nucleotides did not potentiate I_{CatSper} (Supplementary Fig. 3), suggesting that the observed CatSper1-dependent rise in intracellular Ca^{2+} (that is, $[\text{Ca}^{2+}]_i$) induced by cyclic nucleotides⁴ did not result from direct modulation of the CatSper channel. Calcium (2 mM) current through CatSper channels was not sensitive to the L-type Ca_V channels blockers nifedipine ($30 \mu\text{M}$) and nimodipine ($20 \mu\text{M}$). Cadmium ($200 \mu\text{M}$), nickel ($300 \mu\text{M}$) and ruthenium red ($10 \mu\text{M}$) reversibly blocked I_{CatSper} (data not shown).

In summary, we have provided a direct method for reliably reproducible, whole-cell recording of sperm cells. We have directly measured the principal Ca^{2+} -selective current in sperm, shown that it is greatly potentiated by intracellular alkalinization, and demonstrated that it is absent in *Catsper1*^{-/-} mice. To penetrate the cumulus and zona pellucida surrounding the egg, sperm hyperactivate, changing their tail motion from symmetric, fast and low amplitude (sinusoidal) to asymmetric, slow and large amplitude (whiplike)^{1,5}. An increase in $[\text{Ca}^{2+}]_i$ is required for hyperactivation²⁸; CatSper1 is required for mouse sperm hyperactivation⁵ and fertility⁴. Thus, we suggest that during sperm cell capacitation, alkalinization

activates the sperm-specific CatSper Ca^{2+} channel to increase intra-flagellar $[\text{Ca}^{2+}]$ and induce hyperactivated motility.

METHODS

Sperm cell preparation and incubation. Corpus epididymal sperm cells were prepared from male mice aged 3–8 months. Excised corpus epididymides were rinsed in HS medium comprising (in mM): 135 NaCl, 5 KCl, 2 CaCl_2 , 1 MgSO_4 , 20 HEPES, 5 glucose, 10 lactic acid and 1 sodium pyruvate; pH 7.4 (NaOH). Several incisions were made in the epididymis, and sperm cells were shaken from them in HS solution. After 10 min of settling time, the supernatant enriched with highly motile sperm cells was centrifuged, washed, resuspended in HS solution and stored at 25°C for up to 12 h. Beheaded and tail-less sperm cells were prepared by incubating the sperm cell suspension in the presence 0.2 mg ml^{-1} of trypsin at 37°C for 5 min, followed by gentle trituration.

Whole-cell recordings. The patch pipette was applied to the sperm cell cytoplasmic droplet. The cytoplasmic droplet migrates along the sperm cell midpiece from the head to the principal piece during spermatozoa epididymal transport and maturation. In most species the cell sheds it during ejaculation¹⁸. We used spermatozoa from corpus epididymis, because the cytoplasmic droplet of such cells was less fragile. The most mature spermatozoa, those with the cytoplasmic droplet located next to the principal piece, were chosen for experiments. Sperm were placed on untreated glass coverslips. Pipette resistance with caesium methanesulphonate solutions was 3–11 M Ω . Access resistance in the whole-cell configuration was 25–80 M Ω . Roughly five successful break-ins per day on average could be easily achieved. Cells were stimulated every 5 s.

The pipette solutions were made up as follows (in mM). For recordings of monovalent I_{CatSper} at intracellular pH (pH_i) 7.2: 135 caesium methanesulphonate, 5 CsCl, 10 HEPES, 10 EGTA, 5 Na_2ATP and 0.5 Na_2GTP (pH with CsOH); in some experiments, Na_2ATP and Na_2GTP were omitted from the pipette solution without any noticeable effect on I_{CatSper} . For recordings of monovalent I_{CatSper} at intracellular pH 6.0 and pH 8.4: 140 caesium methanesulphonate, 5 CsCl, 20 MES (pH 6.0) or TrisBase (pH 8.4), and 10 EGTA (pH with CsOH). For recordings of divalent currents at pH_i 7.0 and pH_i 8.0: 150 NMDG, 5 CsCl, 10 HEPES, 10 EGTA, 5 Na_2ATP and 0.5 Na_2GTP (pH with methanesulphonic acid); in some experiments, Na_2ATP and Na_2GTP or CsCl were omitted from the pipette solution without any noticeable effect on I_{CatSper} . For recordings of divalent currents at pH_i 6.0 and pH_i 6.5: 165 NMDG, 10 MES and 10 EGTA (pH with methanesulphonic acid). For recordings of divalent currents in symmetrical 50 mM Ba^{2+} conditions: 50 $\text{Ba}(\text{OH})_2$, 90 NMDG and 20 HEPES; pH 8.0 (with methanesulphonic acid). BAPTA solution (175 NMDG, 10 BAPTA, 10 HEPES and 10 MES; pH 6.0 or pH 7.5 with methanesulphonic acid) was used in some experiments in place of EGTA solutions (to eliminate divalent-dependent H^+ release by this chelator).

Seals between the patch pipette and the cytoplasmic droplet were formed in HS bath solution. After break-in, the bath solution could be changed as follows. For monovalent currents in divalent-free conditions (DVF solution): 150 sodium gluconate, 20 HEPES and 5 Na_3HEDTA ; pH 7.4 with NaOH. In some experiments, the DVF solution was modified by substitution of 5 Na_3HEDTA by 2 Na_3HEDTA and 2 EGTA, or by 1 Na_3HEDTA and 1 EGTA. CaCl_2 was added to the '2 Na_3HEDTA and 2 EGTA' DVF solution in accordance with the WinMAXC v2.05 program (C. Patton, Stanford Univ.) to obtain free $[\text{Ca}^{2+}]$ in the nanomolar and low-micromolar ranges. Bath solutions with different concentrations of Ca^{2+} , Ba^{2+} and Mg^{2+} ions were prepared by mixing nominal DVF bath solution (160 NMDG and 20 HEPES; pH 7.4 with methanesulphonic acid) and 50 mM divalent solution ($50 \text{ mM Ca}(\text{OH})_2$, $\text{Ba}(\text{OH})_2$ or $\text{Mg}(\text{OH})_2$, 90 NMDG and 20 HEPES; pH 7.4 with methanesulphonic acid). Osmolarity of the solutions was about 305 mmol kg^{-1} .

Perforated-patch recordings. Pipettes were filled as follows (in mM): 130 caesium methanesulphonate, 8 NaCl, 10 HEPES, 10 Cs_4BAPTA and $240 \mu\text{g ml}^{-1}$ of amphotericin B; pH 7.2. Pipette resistance was 12–17 M Ω . After antibiotic partitioning, access resistance in the whole-cell configuration was 35–100 M Ω . Bath solutions were similar to those used for the conventional whole-cell configuration.

Data analysis. Signals were acquired in PClamp 9 (Axon Instruments). Signals were filtered at 2 kHz and sampled at 10 kHz. The Hill equation was used to fit Fig. 3a, and the Boltzmann equation to fit Fig. 4b (Origin 7.0). Statistical data were calculated as the mean $\pm$ s.e.m.

Received 15 August; accepted 11 November 2005.

- Ho, H. C. & Suarez, S. S. Hyperactivation of mammalian spermatozoa: function and regulation. *Reproduction* **122**, 519–526 (2001).
- Darszon, A., Labarca, P., Nishigaki, T. & Espinosa, F. Ion channels in sperm physiology. *Physiol. Rev.* **79**, 481–510 (1999).

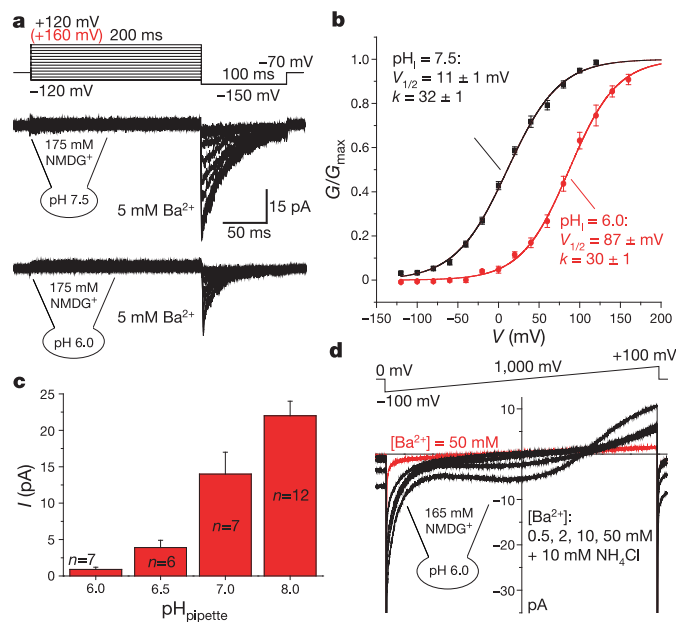


Figure 4 | pH and voltage-dependence of I_{CatSper} . **a**, I_{CatSper} relaxation (tail currents after subtraction of current in $5 \text{ mM } [\text{Mg}^{2+}]_o$) after instantaneous jump to -150 mV from voltages ranging from -120 mV to $+160 \text{ mV}$ ($V_H = -70 \text{ mV}$) at pH_i 7.5 (top) and pH_i 6.0 (bottom). **b**, Activation curves (fit to Boltzmann equation) of I_{CatSper} at pH_i 7.5 and pH_i 6.0, showing that alkalinization increases I_{CatSper} in the physiological range of potential (data are mean $\pm$ s.e.m.). **c**, Histogram showing the dependence of the mean I_{CatSper} Ba^{2+} amplitude on $\text{pH}_{\text{pipette}}$. Ba^{2+} tail currents were elicited by stepping V_H from 0 to -100 mV . Bath $[\text{Ba}^{2+}] = 50 \text{ mM}$; current was measured at -100 mV after subtraction of the capacitive transients. **d**, Addition of $10 \text{ mM NH}_4\text{Cl}$ to the bath solution elicited robust Ba^{2+} I_{CatSper} ($\text{pH}_{\text{pip}} = 6$; $V_H = 0 \text{ mV}$).

3. Darszon, A. *et al.* Calcium channels and Ca^{2+} fluctuations in sperm physiology. *Int. Rev. Cytol.* **243**, 79–172 (2005).
4. Ren, D. *et al.* A sperm ion channel required for sperm motility and male fertility. *Nature* **413**, 603–609 (2001).
5. Carlson, A. E. *et al.* CatSper1 required for evoked Ca^{2+} entry and control of flagellar function in sperm. *Proc. Natl Acad. Sci. USA* **100**, 14864–14868 (2003).
6. Carlson, A. E. *et al.* Identical phenotypes of CatSper1 and CatSper2 null sperm. *J. Biol. Chem.* **280**, 32238–32244 (2005).
7. Quill, T. A., Ren, D., Clapham, D. E. & Garbers, D. L. A voltage-gated ion channel expressed specifically in spermatozoa. *Proc. Natl Acad. Sci. USA* **98**, 12527–12531 (2001).
8. Quill, T. A. *et al.* Hyperactivated sperm motility driven by CatSper2 is required for fertilization. *Proc. Natl Acad. Sci. USA* **100**, 14869–14874 (2003).
9. Babcock, D. F., Bosma, M. M., Battaglia, D. E. & Darszon, A. Early persistent activation of sperm K^{+} channels by the egg peptide speract. *Proc. Natl Acad. Sci. USA* **89**, 6001–6005 (1992).
10. Espinosa, F. *et al.* Mouse sperm patch-clamp recordings reveal single Cl^{-} channels sensitive to niflumic acid, a blocker of the sperm acrosome reaction. *FEBS Lett.* **426**, 47–51 (1998).
11. Gu, Y., Kirkman-Brown, J. C., Korchev, Y., Barratt, C. L. & Publicover, S. J. Multi-state, 4-aminopyridine-sensitive ion channels in human spermatozoa. *Dev. Biol.* **274**, 308–317 (2004).
12. Guerrero, A., Sanchez, J. A. & Darszon, A. Single-channel activity in sea urchin sperm revealed by the patch-clamp technique. *FEBS Lett.* **220**, 295–298 (1987).
13. Sanchez, D., Labarca, P. & Darszon, A. Sea urchin sperm cation-selective channels directly modulated by cAMP. *FEBS Lett.* **503**, 111–115 (2001).
14. Gorelik, J. *et al.* Ion channels in small cells and subcellular structures can be studied with a smart patch-clamp system. *Biophys. J.* **83**, 3296–3303 (2002).
15. Weyand, I. *et al.* Cloning and functional expression of a cyclic-nucleotide-gated channel from mammalian sperm. *Nature* **368**, 859–863 (1994).
16. Florman, H. M., Arnoult, C., Kazam, I. G., Li, C. & O'Toole, C. M. A perspective on the control of mammalian fertilization by egg-activated ion channels in sperm: a tale of two channels. *Biol. Reprod.* **59**, 12–16 (1998).
17. Lobley, A., Pierron, V., Reynolds, L., Allen, L. & Michalovich, D. Identification of human and mouse CatSper3 and CatSper4 genes: characterisation of a common interaction domain and evidence for expression in testis. *Reprod. Biol. Endocrinol.* **1**, 53 (2003).
18. Cooper, T. G. Cytoplasmic droplets: the good, the bad or just confusing? *Hum. Reprod.* **20**, 9–11 (2005).
19. Gonzalez-Martinez, M. T. Induction of a sodium-dependent depolarization by external calcium removal in human sperm. *J. Biol. Chem.* **278**, 36304–36310 (2003).
20. Espinosa, F. & Darszon, A. Mouse sperm membrane potential: changes induced by Ca^{2+} . *FEBS Lett.* **372**, 119–125 (1995).
21. Horn, R. & Marty, A. Muscarinic activation of ionic currents measured by a new whole-cell recording method. *J. Gen. Physiol.* **92**, 145–159 (1988).
22. Hille, B. *Ion Channels of Excitable Membranes* (Sinauer Associates, Sunderland, MA, 2001).
23. Sather, W. A. & McCleskey, E. W. Permeation and selectivity in calcium channels. *Annu. Rev. Physiol.* **65**, 133–159 (2003).
24. Fraire-Zamora, J. J. & Gonzalez-Martinez, M. T. Effect of intracellular pH on depolarization-evoked calcium influx in human sperm. *Am. J. Physiol. Cell. Physiol.* **287**, C1688–C1696 (2004).
25. Santi, C. M., Santos, T., Hernandez-Cruz, A. & Darszon, A. Properties of a novel pH-dependent Ca^{2+} permeation pathway present in male germ cells with possible roles in spermatogenesis and mature sperm function. *J. Gen. Physiol.* **112**, 33–53 (1998).
26. Babcock, D. F. Examination of the intracellular ionic environment and of ionophore action by null point measurements employing the fluorescein chromophore. *J. Biol. Chem.* **258**, 6380–6389 (1983).
27. Zeng, Y., Oberdorf, J. A. & Florman, H. M. pH regulation in mouse sperm: identification of Na^{+} , Cl^{-} , and HCO_3^{-} -dependent and arylaminobenzoate-dependent regulatory mechanisms and characterization of their roles in sperm capacitation. *Dev. Biol.* **173**, 510–520 (1996).
28. Ho, H. C., Granish, K. A. & Suarez, S. S. Hyperactivated motility of bull sperm is triggered at the axoneme by Ca^{2+} and not cAMP. *Dev. Biol.* **250**, 208–217 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Qi for help in preparation of sperm cells and for discussion.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to D.E.C. (dclapham@enders.tch.harvard.edu).

Structural basis for Duffy recognition by the malaria parasite Duffy-binding-like domain

Saurabh Kumar Singh¹, Rachna Hora¹, Hassan Belrhali³, Chetan E. Chitnis² & Amit Sharma¹

Molecular processes that govern pathogenic features of erythrocyte invasion and cytoadherence in malaria are reliant on *Plasmodium*-specific Duffy-binding-like domains (DBLs)¹. These cysteine-rich modules recognize diverse host cell-surface receptors during pathogenesis. DBLs of parasite erythrocyte-binding proteins mediate invasion, and those from the antigenically variant *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) have been implicated in cytoadherence. The simian and human malarial parasites, *P. knowlesi* and *P. vivax*, invade human erythrocytes exclusively through the host DARC receptor (Duffy antigen receptor for chemokines)^{2,3}. Here we present the crystal structure of the *P. knowlesi* DBL domain (Pk α -DBL), which binds to DARC during invasion of human erythrocytes. Pk α -DBL retains the overall fold observed in DBLs from *P. falciparum* erythrocyte-binding antigen (EBA)-175 (ref. 4). Mapping the residues that have previously been implicated in binding^{5–7} highlights a fairly flat but exposed site for DARC recognition in subdomain 2 of Pk α -DBL; this is in sharp contrast to receptor recognition by EBA-175 (ref. 4). In Pk α -DBL, the residues that contact DARC and the clusters of residues under immune pressure map to opposite surfaces of the DBL, and suggest a possible mechanism for immune evasion by *P. vivax*. Our comparative structural analysis of Pk α -DBL and *P. falciparum* EBA-175 provides a framework for the understanding of malaria parasite DBLs, and may affect the development of new prophylactic and therapeutic strategies.

Plasmodium parasites are the causative agents of malaria, which affects more than 500 million people and kills approximately two million people annually⁸. Although invasion of erythrocytes drives the common symptoms of malaria infection, erythrocyte

cytoadherence is also implicated in the progression to severe disease. The DBLs from *P. vivax* and *P. knowlesi* erythrocyte-binding proteins (Pv-DBL and Pk α -DBL, respectively) are very similar in sequence (71% identity) and are homologous to the two DBLs (F1 and F2) in *P. falciparum* EBA-175, which target sialic acid residues on human erythrocyte glycoprotein A for invasion¹. A key component of the interaction between human DARC and Pv-DBL or Pk α -DBL is a sulphated tyrosine (Tyr 41) residue on DARC⁹; this is reminiscent of the role of sulphated tyrosine in the gp120–CCR5 interaction during HIV invasion¹⁰.

Here we determined the crystal structure of recombinant Pk α -DBL⁵ (337 residues) by collecting single-wavelength anomalous dispersion data from a co-crystal of Pk α -DBL and potassium gold cyanide (Table 1). Pk α -DBL crystals were obtained in the presence of the detergent *N*-octyl β -D-glucopyranoside, which was required for all crystal-related manipulations. We calculated the initial electron density map to 3.3 Å resolution, and data from a better-diffracting native crystal allowed model refinement to 3 Å.

Pk α -DBL is an all-helical, boomerang-shaped, monomeric module containing twelve helices spread over three distinct subdomains (Fig. 1). Despite poor sequence conservation, the overall structure of Pk α -DBL is similar to the F1/F2 DBLs from EBA-175 (ref. 4). In Pk α -DBL, a random-coil stretch of residues 15–52 constitutes subdomain 1, and residues 64–180 (α 1– α 6) make up subdomain 2. A short linker segment attaches subdomain 2 to subdomain 3, and subdomain 3 extends from residues 186–307 and encompasses helices α 7– α 12. The segment between subdomains 1 and 2 (residues 53–63) is missing and is likely to be disordered, consistent with our earlier proteolysis data⁵, which predicted a flexible linker starting at

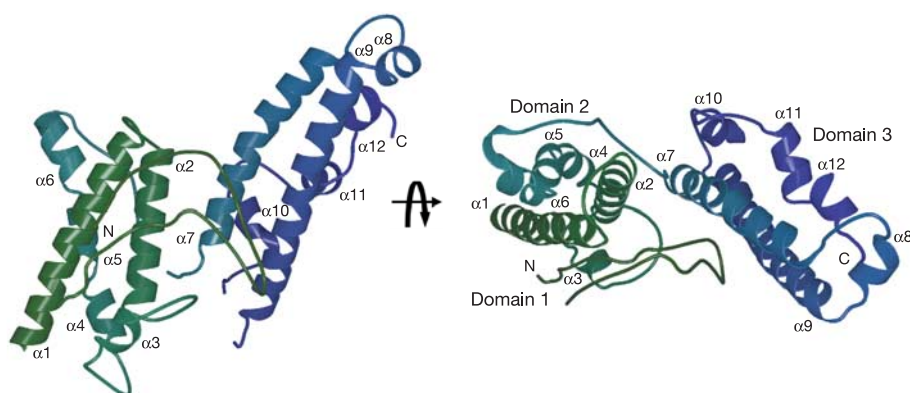


Figure 1 | Structure of Pk α -DBL. Orthogonal views of the DBL domain. Pk α -DBL has a compact, flat and multi-subdomain architecture. The three subdomains of Pk α -DBL are subdomain 1 (residues 15–52), subdomain 2

(residues 64–180) and subdomain 3 (residues 186–307). Note the intermediate, one-sided placement of subdomain 1, which is connected to subdomain 2 by a proteolysis-sensitive linker⁵.

¹Structural and Computational Biology Group and ²Malaria Research Group, International Centre for Genetic Engineering and Biotechnology (ICGEB), Aruna Asaf Ali Marg, New Delhi-110067, India. ³European Molecular Biology Laboratory (EMBL), 6 rue Jules Horowitz, BP 181, F-38042 Grenoble Cédex 9, France.

Table 1 | Diffraction data collection and model refinement statistics

	Native	Gold-SAD
Wavelength (Å)	0.88560	1.03927
Range (Å)	20–3.0	50–3.3
Reflections	9,599	7,600*
R_{merge} (%)†	7.5 (48.3)	9.5 (37.2)
Completeness (%)	98.7 (97.6)	100 (100)

* Keeping Bijvoet pairs separate during data processing.
† $R_{\text{merge}} = \sum (I - \langle I \rangle) / \sum \langle I \rangle$; numbers in parentheses correspond to values in the highest resolution shell.

residue Arg 57. Subdomain 3 consists of two ~40 Å long helical towers (α7 and α9) that are connected by a surface-exposed linker we term the subdomain 3 loop. This segment is strapped onto subdomain 3 and has three disulphide bridges in its vicinity.

Notably, subdomain 1, which is not required for the Pkα-DBL–DARC interaction⁵, rests between subdomains 2 and 3 on one side of the DBL domain and makes hydrophobic contacts with both subdomains. Pkα-DBL has no inter-subdomain disulphide bridges, and the three subdomains remain connected by sharing hydrophobic surfaces and flexible linkers. Subdomains 1, 2 and 3 contain two, one and three disulphide bridges, respectively. These disulphide bridges are a hallmark of DBLs and are superimposable between F1, F2 and Pkα-DBL (Fig. 2a, b and Supplementary Fig. 1). F1 and Pkα-DBL each contain six identically linked disulphide bridges. F2 contains seven disulphide bridges (of which six are conserved between F1, F2 and Pkα-DBL), with the extra bridge present in subdomain 3. Despite overall structural conservation between F1–F2 (the two domains linked in tandem) and Pkα-DBL, their oligomeric states differ significantly. Recombinant F1–F2 shows concentration-dependent dimerization⁴, whereas Pv-DBL and Pkα-DBL are monomeric in nature. This difference is also evident from sequence alignments (Fig. 2b), which show that the F1–F2 dimerization elements are absent from Pv-DBL and Pkα-DBL, and are not conserved in DBLs from PfEMP1 proteins.

Structural features in Pv-DBL and Pkα-DBL that are required for binding human erythrocytes have previously been detailed in domain deletion⁵ and site-directed mutagenesis experiments^{6,7}. On the basis of these data, we identify an area on subdomain 2 that provides the essential, invariant and surface-exposed contact residues Tyr 94, Asn 95, Lys 96, Arg 103, Leu 168 and Ile 175 required for recognition of DARC on human erythrocytes (Fig. 3a). Site-directed mutagenesis of each of these residues, which are identical between Pkα-DBL and Pv-DBL (Fig. 2b), abrogates or disrupts the Pv/Pkα-DBL–DARC interaction^{6,7}. These residues project on to a relatively flat, dual-character surface on the top of subdomain 2 (Fig. 3a, b) that contains distinct nonpolar (Tyr 94, Leu 168 and Ile 175) and polar residues (Asn 95, Lys 96 and Arg 103). The nonpolar components of the DARC-recognition site lie in apposition to a cluster of basic residues that includes Lys 96, Lys 100, Arg 103 and Lys 177

Table 2 | X-ray refinement statistics

Resolution (Å)	20.0–3.0
$R_{\text{work}}/R_{\text{free}}$	24.8/31.2
Number of atoms	
Protein	2,186
Ligand/ion	0
Water	42
B-factors	
Protein	71
Ligand/ion	0
Water	43
r.m.s deviations	
Bond length (Å)	0.014
Bond angles (°)	1.9

* $R_{\text{work}} = \sum (|F_{\text{obs}}| - |F_{\text{calc}}|) / \sum |F_{\text{obs}}|$. R_{free} is the same as R_{work} but comprises a 'test' set of reflections (fraction 6%; number 556) that were excluded during model refinement³⁰.

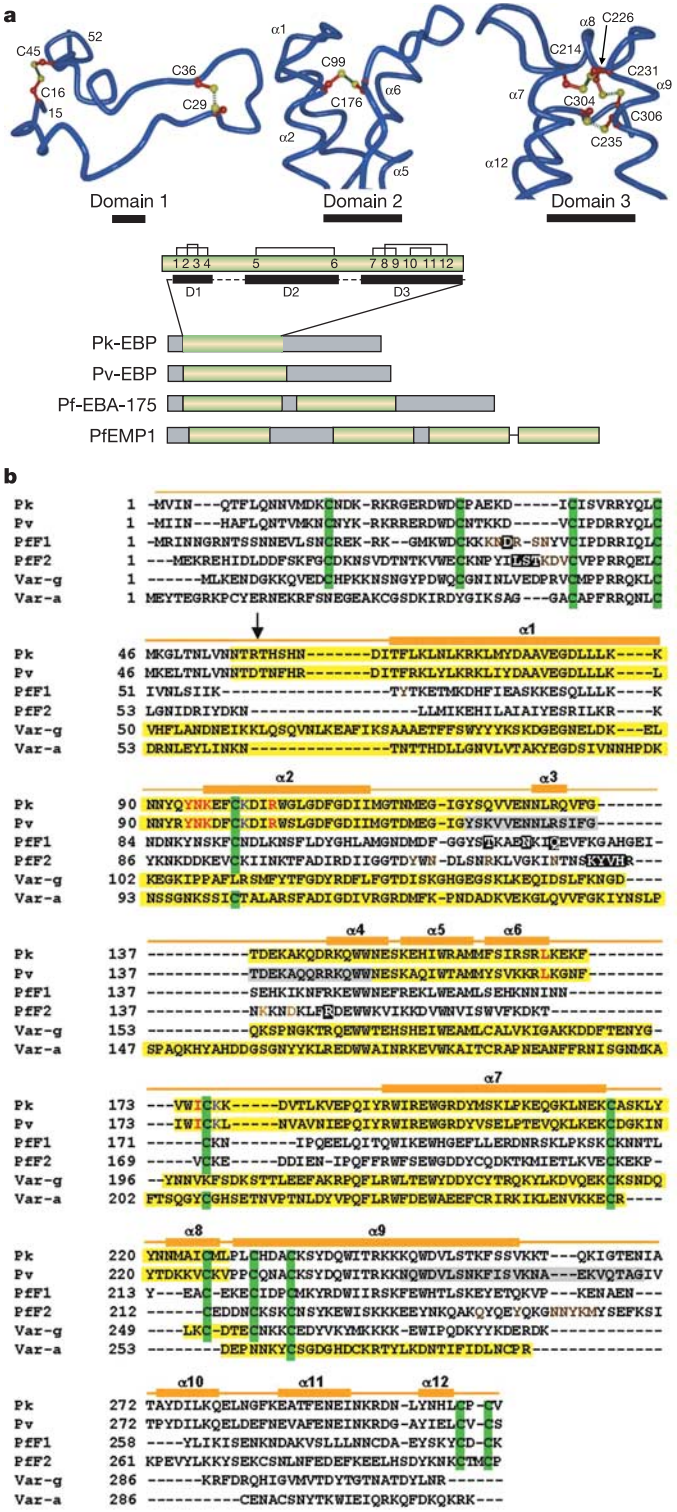


Figure 2 | Disulphide bond distribution. **a**, The top panel shows the disulphide bonds present in subdomains 1, 2 and 3 of Pkα-DBL. The pattern of disulphide bridges is also shown on a linear domain map (green box) in the lower panel. Homologues of Pkα-DBL (green boxes) are present in many EBP and PfEMP1 proteins, the extracellular domains of which are shown in the lower panel. **b**, Sequence alignment of homologous DBLs from three EBPs and two PfEMP1 proteins for which receptor-binding regions have been mapped¹⁸. Shown are secondary structure elements for Pkα-DBL/Pv-DBL (orange), minimal receptor-binding regions (yellow)¹⁸, conserved cysteine residues (green), proteolytic site on Pkα-DBL (arrow), Pv-DBL clusters of polymorphic residues (grey), mutations that affect DARC recognition (red), contact residues between F1–F2 and sialic acid (brown) and F1–F2 dimerization elements (black boxes)⁴.

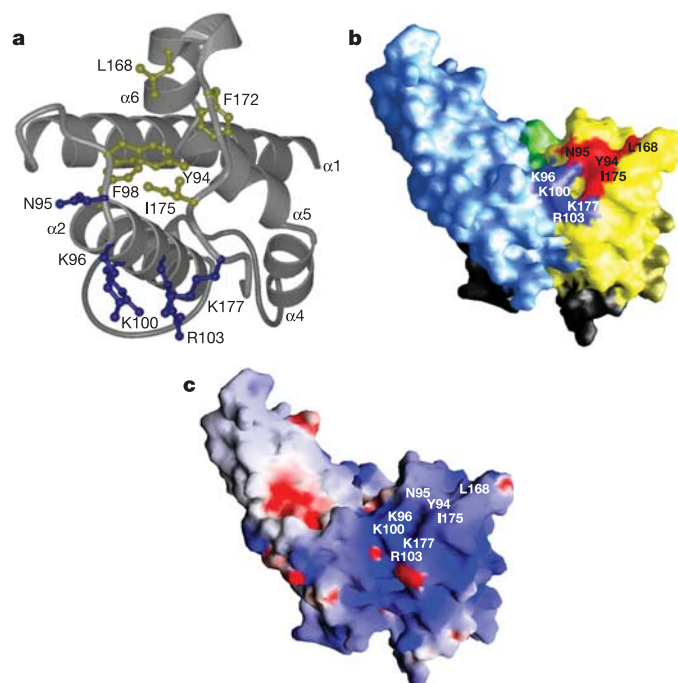


Figure 3 | DARC recognition site. **a**, Top-down view of subdomain 2 (grey) with DARC-contact residues shown in ball-and-stick representation. Note the distinct dual character of the DARC-recognition site, which includes nonpolar (yellow) and polar residues (blue). **b**, Molecular surface representation of Pk α -DBL, showing subdomain 1 (green), subdomain 2 (yellow) and subdomain 3 (light blue), amino acid residues for which mutation disrupts recognition of DARC (red), basic residues (blue) and clusters of polymorphic residues (grey). **c**, GRASP²⁹ representation of Pk α -DBL electropotential surface with positive (blue) and negative (red) patches. Residues proximal to the DARC-binding site (Lys 96, Lys 100, Arg 103 and Lys 177) form a continuous patch of electropositive density (blue).

(Fig. 3b, c); these basic residues may interact with the sulphate group of DARC Tyr 41.

Sulphation of Tyr 41 is known to increase Pv/Pk α -DBL–DARC affinity by up to 1,000-fold, suggesting that the charge characteristics of sulfated tyrosine on DARC and its binding site on Pv/Pk α -DBL are central to successful docking. Consistent with our description of the DARC-interaction site on Pk α -DBL, several single-site mutants that do not affect DARC binding fall outside the recognition surface^{6,7}. Other mutants that decrease DARC affinity for Pv-DBL, such as Phe 98 and Phe 172, are likely to have purely structural roles because of their proximity to the binding site. Our analysis suggests that DARC recognition by Pv-DBL/Pk α -DBL is achieved via a single, contiguous and fully exposed site on subdomain 2. In contrast, the F1-F2 dimer contains six sialic acid binding sites that are scattered at the dimer interface, of which four are in channels and two are on the F1-F2 surface⁴. Engagement of sialic acid/glycophorin A by dimeric F1-F2 would therefore necessitate significant conformational or oligomeric rearrangements of interacting molecules on parasite and host cell surfaces. In contrast, the monomeric nature of Pv-DBL/Pk α -DBL and its solvent-exposed binding site for DARC obviates any requirement for gross conformational changes during DARC engagement.

Subdomain 3 of Pk α -DBL is a principal structural feature, and its underlying sequences are conserved across DBLs (Fig. 2b). The subdomain 3 loop (Ala 215 to Asn 221) juts out into the solvent and rests on top of helical towers α 7 and α 9. Among *Plasmodium* species, both the sequence and length of this loop vary greatly (Fig. 2b). The subdomain 3 loop is adjacent to the three highly conserved disulphide bridges, where the Cys 7–Cys 9 and Cys 8–Cys 12 pairs pin this loop on top of α 7 and α 9, and orient it such that the

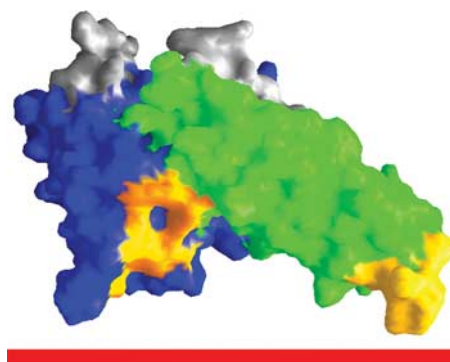


Figure 4 | Model of Pv-DBL-erythrocyte interaction. Molecular surface showing Pv-DBL clusters of polymorphic residues (grey) and the DARC-recognition site (orange). Subdomains 1, 2 (blue) and 3 (green) are shown. Note the position of the subdomain 3 loop (yellow). Docking of subdomain 2 onto DARC would bring the subdomain 3 loop close to the host erythrocyte membrane (red rectangle).

interaction of DARC with the binding site in subdomain 2 will necessarily bring this segment in proximity to the host membrane. The functional significance of either subdomain 3 or its prominent loop remains unclear.

Mapping of the Pv-DBL amino acids shown to be polymorphic among *P. vivax* field isolates from Papua New Guinea and Colombia^{11–13} is revealing (Fig. 3b). The DARC-recognition site and clusters of polymorphic residues lie on opposite sides, and this spatial segregation corroborates the ‘just-in-time’ release of Pv-EBP/Pk α -EBP from micronemes^{7,14}. Secretion of Pv-EBP from micronemes is probably followed by rapid DARC engagement, burying the binding site residues and enabling immune response evasion. According to this model (Fig. 4), residues coming under immune attack will be on the opposite surface, which is where polymorphism is predominantly manifested—we suggest this might be to escape immune clearance. This parasite strategy for evading neutralization is different from HIV, for example, in which the receptor-binding residues on gp120 are obscured by glycosylation/conformational camouflaging^{15,16}, and from influenza haemagglutinin, in which sequence polymorphisms tend to surround the active site pocket¹⁷.

Notably, many malaria parasite species use DBLs as molecular adaptors for processes related to erythrocyte invasion⁴ and cytoadherence¹⁸. Comparison of F1-F2 with Pk α -DBL suggests that the sequence variability in DBLs underpins the functional diversity that manifests itself in varied receptor-recognition abilities. The interaction between EBA-175 and sialic acid on glycophorin A is non-essential, owing to redundant *P. falciparum* erythrocyte invasion pathways¹⁹. Therefore, in order to block multiple *P. falciparum* invasion routes, EBA-175 and its homologues must be targeted simultaneously. In contrast, the obligatory nature of the Pv-DBL–DARC interaction for *P. vivax* invasion of human erythrocytes underscores its significance for therapeutic approaches. The structural analysis of Pv-DBL/Pk α -DBL provides a platform for targeting the exposed DARC-binding site for development of specific inhibitors or inhibitory antibodies.

Our observation that naturally occurring sequence polymorphisms in Pv-DBL cluster on surfaces distant from the DARC-binding site bodes well for the use of Pv-DBL as a vaccine against *P. vivax* malaria: antibodies induced by Pv-DBL administration that are capable of blocking the Pv-DBL–DARC interaction are unlikely to be affected by polymorphisms in different field strains of *P. vivax*. The structure of Pk α -DBL may therefore enable the development of strategies to inhibit *P. vivax* invasion of human erythrocytes.

METHODS

Protein crystallization and data collection. Recombinant Pk α -DBL (337 residues) was produced according to published protocols⁵. This protein binds

to human erythrocytes⁵ and to the sulphated form (Tyr 41) of human DARC with specificity (Supplementary Fig. 2). Purified Pk α -DBL (~15 mg ml⁻¹) was crystallized in 0.1 M HEPES pH 7.5, 12% polyethylene glycol 10,000, 0.1 M ammonium sulphate and 0.5% *N*-octyl β -D-glucopyranoside. The crystal used for single-wavelength anomalous dispersion (SAD) data collection was grown in 10 mM potassium aurocyanide with 2% v/v dioxane, 0.1 M bicine pH 9.0 and 10% polyethylene glycol 20,000. All crystal forms belong to the space group *P*₄₃₂₁2 with *a* = 60.57 Å, *b* = 60.57 Å, *c* = 241.04 Å and with one monomer per asymmetric unit. Crystals were soaked in mother liquor supplemented with 20% ethylene glycol for 30 s and then flash-frozen before data collection. The native and heavy atom co-crystals were extremely sensitive to both temperature changes and X-ray radiation, and diffracted poorly at in-house sources. All diffraction data were collected using synchrotron radiation under cryogenic conditions at the BM14 beamline, ESRF, Grenoble. Relevant data collection and refinement statistics are provided in Table 1.

Structure determination and analysis. Diffraction data were collected on a ~50 × 50 × 100 μm protein crystal of Pk α -DBL co-crystallized with gold. The Au-absorption edge was determined by a fluorescence scan and 360° of data were collected in 1.0° oscillation frames using a MARCCD detector (Table 1). Native diffraction data to 3.0 Å were also collected at the BM14 beamline. The data were indexed and scaled with DENZO and SCALEPACK programs²⁰. Processing of the SAD data set was done using SHARP²¹ and SOLVE²², and each program identified two gold sites. SAD phases, after density modification (solvent content of the crystals is 62%), were used to calculate a native map to 3.3 Å, which was of moderate quality although the secondary structure elements (helices) were clearly visible. Backbone and side-chain tracing of the protein structure was done using the program O²³. Crystallographic model refinement was initiated using SAD data in CNS²⁴, and the model was subsequently refined against native data to 3 Å. A total of 6% (or 556) of the reflections were kept aside for *R*_{free} evaluation before start of the refinement.

The final model statistics are given in Table 2; there are no residues in disallowed regions of the Ramachandran plot and the model stereochemistry is excellent, as verified by PROCHECK²⁵. Where side chain positions are not clear, some residues in the final model have been modelled as alanines. In addition, no reliable model could be built for residues 1–14, 53–63, 181–185, 263–268 and 308–337. Structural superpositions between F1/F2 (PDB accession code 1ZRL) and Pk α -DBL were done using LSQMAN²⁶, and the residue limits along with root mean square deviations are F1(8–282)–Pk α -DBL(15–307), 1.6 Å (for 195 C α atoms); F1(8–282)–F2(297–596), 1.9 Å (for 185 C α atoms), and F2(297–596)–Pk α -DBL(15–307), 1.8 Å (for 156 C α atoms). All figures were made using BOBSCRIPT²⁷, RASTER3D²⁸ or GRASP²⁹.

Received 15 September; accepted 14 November 2005.

Published online 21 December 2005.

1. Miller, L. H., Baruch, D. I., Marsh, K. & Doumbo, O. K. The pathogenic basis of malaria. *Nature* **415**, 673–679 (2002).
2. Miller, L. H., Mason, S. J., Dvorak, J. A., McGinniss, M. H. & Rothman, I. K. Erythrocyte receptors for (*Plasmodium knowlesi*) malaria: Duffy blood group determinants. *Science* **189**, 561–563 (1975).
3. Miller, L. H., Mason, S. J., Clyde, D. F. & McGinniss, M. H. The resistance factor to *Plasmodium vivax* in blacks. The Duffy-blood-group genotype, FyFy. *N. Engl. J. Med.* **295**, 302–304 (1976).
4. Tolia, N. H., Enemark, E. J., Sim, B. K. & Joshua-Tor, L. Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite *Plasmodium falciparum*. *Cell* **122**, 183–193 (2005).
5. Singh, S. K. *et al.* Definition of structural elements in *Plasmodium vivax* and *P. knowlesi* Duffy-binding domains necessary for erythrocyte invasion. *Biochem. J.* **374**, 193–198 (2003).
6. Hans, D. *et al.* Mapping binding residues in the *Plasmodium vivax* domain that binds Duffy antigen during red cell invasion. *Mol. Microbiol.* **55**, 1423–1434 (2005).
7. VanBuskirk, K. M., Sevova, E. & Adams, J. H. Conserved residues in the *Plasmodium vivax* Duffy-binding protein ligand domain are critical for erythrocyte receptor recognition. *Proc. Natl Acad. Sci. USA* **101**, 15754–15759 (2004).
8. Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y. & Hay, S. I. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature* **434**, 214–217 (2005).
9. Choe, H. *et al.* Sulphated tyrosines mediate association of chemokines and *Plasmodium vivax* Duffy binding protein with the Duffy antigen/receptor for chemokines (DARC). *Mol. Microbiol.* **55**, 1413–1422 (2005).
10. Farzan, M. *et al.* Tyrosine sulfation of the amino terminus of CCR5 facilitates HIV-1 entry. *Cell* **96**, 667–676 (1999).
11. Tsuboi, T. *et al.* Natural variation within the principal adhesion domain of the *Plasmodium vivax* Duffy binding protein. *Infect. Immun.* **62**, 5581–5586 (1994).
12. Xainli, J., Adams, J. H. & King, C. L. The erythrocyte binding motif of *Plasmodium vivax* Duffy binding protein is highly polymorphic and functionally conserved in isolates from Papua New Guinea. *Mol. Biochem. Parasitol.* **111**, 253–260 (2000).
13. Ampudia, E., Patarroyo, M. A., Patarroyo, M. E. & Murillo, L. A. Genetic polymorphism of the Duffy receptor binding domain of *Plasmodium vivax* in Colombian wild isolates. *Mol. Biochem. Parasitol.* **78**, 269–272 (1996).
14. Adams, J. H. *et al.* The Duffy receptor family of *Plasmodium knowlesi* is located within the micronemes of invasive malaria merozoites. *Cell* **63**, 141–153 (1990).
15. Wyatt, R. *et al.* The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature* **393**, 705–711 (1998).
16. Kwong, P. D. *et al.* HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature* **420**, 678–682 (2002).
17. Wilson, I. A. & Cox, N. J. Structural basis of immune recognition of influenza virus hemagglutinin. *Annu. Rev. Immunol.* **8**, 737–771 (1990).
18. Mayor, A. *et al.* Receptor-binding residues lie in central regions of Duffy-binding-like domains involved in red cell invasion and cytoadherence by malaria parasites. *Blood* **105**, 2557–2563 (2005).
19. Reed, M. B. *et al.* Targeted disruption of an erythrocyte binding antigen in *Plasmodium falciparum* is associated with a switch toward a sialic acid-independent pathway of invasion. *Proc. Natl Acad. Sci. USA* **97**, 7509–7514 (2000).
20. Otwinowski, Z. & Minor, W. Processing of X-ray data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
21. de La Fortelle, E. & Bricogne, G. Maximum-likelihood heavy-atom parameter refinement in the MIR and MAD methods. *Methods Enzymol.* **276**, 472–494 (1997).
22. Terwilliger, T. C. & Berendzen, J. Automated MAD and MIR structure solution. *Acta Crystallogr. D* **55**, 849–861 (1999).
23. Jones, T. A., Zou, J. Y. & Cowan, S. W. Kjeldgaard. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
24. Brunger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
25. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check stereochemical quality of protein structure. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
26. Kleywegt, G. J. Experimental assessment of differences between related protein crystal structures. *Acta Crystallogr. D* **55**, 1878–1884 (1999).
27. Esnouf, R. M. An extensively modified version of MolScript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model.* **15**, 132–134, 112–113 (1997).
28. Merritt, E. A. & Bacon, D. J. Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277**, 505–524 (1997).
29. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).
30. Brunger, A. T. The free *R* value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–474 (1992).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Stuart and M. Walsh for facilitating data collection. We also thank past and present members of ICGEB, A. Rushdi Shakri and S. Krishnaswamy for help. C.E.C. is a Howard Hughes International Research Scholar. The Sharma laboratory is supported by a grant from the European Commission FP6. C.E.C. and A.S. are International Wellcome Trust Senior Research Fellows.

Author Information Atomic coordinates for Pk α -DBL have been deposited in the Protein Data Bank under accession number 2c6j. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.S. (asharma@icgeb.res.in).

Antiviral treatment is more effective than smallpox vaccination upon lethal monkeypox virus infection

Koert J. Stittelaar¹, Johan Neyts³, Lieve Naesens³, Geert van Amerongen^{1,4}, Rob F. van Lavieren⁵, Antonin Holý⁶, Erik De Clercq³, Hubert G. M. Niesters¹, Edwin Fries¹, Chantal Maas¹, Paul G. H. Mulder², Ben A. M. van der Zeijst⁴ & Albert D. M. E. Osterhaus¹

There is concern that variola virus, the aetiological agent of smallpox, may be used as a biological weapon. For this reason several countries are now stockpiling (vaccinia virus-based) smallpox vaccine. Although the preventive use of smallpox vaccination has been well documented, little is known about its efficacy when used after exposure to the virus. Here we compare the effectiveness of (1) post-exposure smallpox vaccination and (2) antiviral treatment with either cidofovir (also called HPMPC or Vistide) or with a related acyclic nucleoside phosphonate analogue (HPMPO-DAPy) after lethal intratracheal infection of cynomolgus monkeys (*Macaca fascicularis*) with monkeypox virus (MPXV). MPXV causes a disease similar to human smallpox¹ and this animal model can be used to measure differences in the protective efficacies of classical and new-generation candidate smallpox vaccines². We show that initiation of antiviral treatment 24 h after lethal intratracheal MPXV infection, using either of the antiviral agents and applying various systemic treatment regimens, resulted in significantly reduced mortality and reduced numbers of cutaneous monkeypox lesions. In contrast, when monkeys were vaccinated 24 h after MPXV infection, using a standard human dose of a currently recommended smallpox vaccine (Elstree-RIVM), no significant reduction in mortality was observed. When antiviral therapy was terminated 13 days after infection, all surviving animals had virus-specific serum antibodies and antiviral T lymphocytes. These data show that adequate preparedness for a biological threat involving smallpox should include the possibility of treating exposed individuals with antiviral compounds such as cidofovir or other selective anti-poxvirus drugs.

The eradication of variola virus by a worldwide vaccination campaign led by the World Health Organization (WHO) in the 1970s (ref. 3), together with the subsequent discontinuation of mass smallpox vaccination with vaccinia virus, have resulted in low or non-existent smallpox immunity in today's human population⁴. The possible introduction of variola virus or MPXV as a biological weapon, or the zoonotic spread of MPXV^{3,5}, is of growing concern. Given the relatively high rate of severe side effects associated with classical smallpox vaccination, reintroduction of general preventive smallpox vaccination has not been recommended⁶. Current plans do not envisage mass vaccination with classical smallpox vaccines until after an outbreak has been detected.

Although the effectiveness of classical smallpox vaccination when used preventively has been well documented, and post-exposure vaccination is part of the WHO eradication strategy, little is known about its efficacy when used after exposure to variola virus. Therefore, treatment with antiviral drugs such as cidofovir or related

compounds that have long-lasting antiviral activity should be considered as an alternative or additional strategy for the prevention or post-exposure treatment of smallpox.

Cidofovir, a nucleotide analogue that is licensed for the treatment of cytomegalovirus retinitis in patients with HIV infection, has been shown to be active against a number of poxviruses, including vaccinia-, variola-, monkeypox- and cowpox virus, and has shown efficacy in immunodeficient patients with severe molluscum contagiosum virus and orf virus infections⁷. Concern that the virulence of variola virus and other poxviruses could be increased by genetic modification would be another reason for developing antiviral

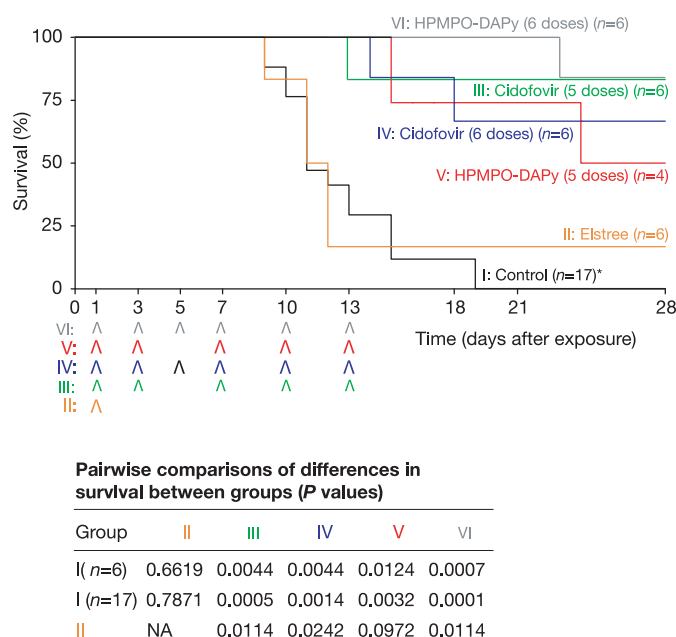


Figure 1 | Kaplan-Meier plot showing the effect of different post-exposure treatments on the survival of monkeys. Survival expressed as percentages after intratracheal MPXV infection. After virus exposure, macaques were sham-treated (group I), vaccinated with smallpox vaccine Elstree-RIVM (group II), treated with 5 doses of cidofovir (group III), with 6 doses of cidofovir (group IV), with 5 doses of HPMPO-DAPy (group V) or with 6 doses of HPMPO-DAPy (group VI). Arrowheads indicate times of vaccination and antiviral treatment. *The control group is supplemented with 11 animals from previous experiments. Statistical analysis of survival rates between groups is shown. NA, not applicable.

¹Department of Virology and ²Department of Epidemiology and Biostatistics, Erasmus MC, 3000 DR Rotterdam, The Netherlands. ³Rega Institute for Medical Research, K.U. Leuven, B-3000 Leuven, Belgium. ⁴Netherlands Vaccine Institute, 3720 AL Bilthoven, The Netherlands. ⁵ViroClinics B.V., 3000 DR Rotterdam, The Netherlands. ⁶Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, 16610 Prague 6, Czech Republic.

strategies against poxvirus infections in humans^{3,8,9}. In several mouse models for poxvirus infections, it has been shown that cidofovir treatment may delay or reduce mortality^{10–16} (see refs 17–19 for a review). Although use of cidofovir has been proposed for treatment of human smallpox²⁰, its effectiveness against orthopoxvirus infections in primates, including humans, has not been addressed to date. Although other anti-poxvirus compounds with different modes of anti-poxvirus activity have been described, their clinical use is not expected in the near future^{21–24}.

Here we compare the efficacy of (1) smallpox vaccination with the classical vaccinia virus-based smallpox vaccine currently stockpiled in The Netherlands (Elstree-RIVM) and (2) antiviral treatment with either cidofovir or a related experimental analogue (HPMPO-DAPy)^{25,26}, against a lethal respiratory MPXV infection of cynomolgus monkeys (*Macaca fascicularis*). This is the most prevalent natural route of infection and the model is considered to be an appropriate surrogate for human smallpox, which meets regulatory requirements outlined by the so-called ‘animal rule’²⁷. It has been successfully used for comparing the efficacy and safety of classical and new-generation candidate smallpox vaccines, including Elstree-RIVM². Indeed, preventive Elstree-RIVM vaccination was highly effective against this challenge². Historical and anecdotal evidence suggests that classical smallpox vaccination may be successful up to 4–5 days after exposure²⁸. This ‘window of opportunity’ is probably related to the relatively long incubation time (7–14 days) for smallpox under natural conditions with relatively low infection doses²⁸. On the basis of body temperature measurement, the incubation time for the MPXV-macaque model is about three days, and is probably due to the high lethal respiratory dose of MPXV given directly into the respiratory tract. This may mimic exposure of the first human individuals during a bioterrorist attack².

We infected six groups of macaques with a lethal respiratory infection dose of 10^7 plaque-forming units (p.f.u.) MPXV per animal, and initiated different antiviral intervention strategies 24 h later. Group I ($n = 6$) served as an untreated control. Group II ($n = 6$) was vaccinated intracutaneously with a human dose of Elstree-RIVM using a bifurcated needle. Group III ($n = 6$) and group IV ($n = 6$) were given 5 mg kg^{-1} cidofovir intraperitoneally (i.p.) on days 1, 3, 7, 10 and 13 after infection or on days 1, 3, 5, 7, 10 and 13 after infection, respectively. Group V ($n = 4$) and group VI ($n = 6$) were given 5 mg kg^{-1} HPMPO-DAPy i.p. on days 1, 3, 7, 10 and 13 after infection or on days 1, 3, 5, 7, 10 and 13 after infection, respectively. As the pharmacokinetics of both antiviral compounds in monkeys proved to be largely similar to those of cidofovir in humans (see Supplementary Information), we chose antiviral doses of cidofovir and HPMPO-DAPy equivalent to those recommended for cidofovir in humans. All monkeys that received antiviral treatment also received supportive treatment with 25 mg kg^{-1} probenecid, administered orally. Cidofovir should always be given under the cover of probenecid, a competitive inhibitor of organic anion transport, to prevent cidofovir nephrotoxicity²⁹. Probenecid should be used with caution in children, pregnant women and persons with sulfa drug allergy. In addition, monkeys received 30 ml of Ringer’s solution i.p. when antiviral treatment was given, to improve hydration.

All six animals in the control group died within 15 days of infection. For statistical comparisons of mortality rates between the groups, 11 control animals from previous experiments were included in the analyses, resulting in a mean day of death of 12.6 days (range 9–19) (Fig. 1). From a total of 34 animals, 17 died as a result of MPXV infection (Fig. 1). Severe morbidity, extensive cutaneous monkeypox lesions, dyspnea and low blood oxygen (O_2) saturation were observed in all control animals (group I; Fig. 2 and Supplementary Information). Vaccination with RIVM-Elstree (group II) protected one animal from MPXV-related death (Figs 1, 2) and severe morbidity. In contrast, cidofovir treatment protected five animals in group III (five-dose cidofovir regimen) and four animals in group IV

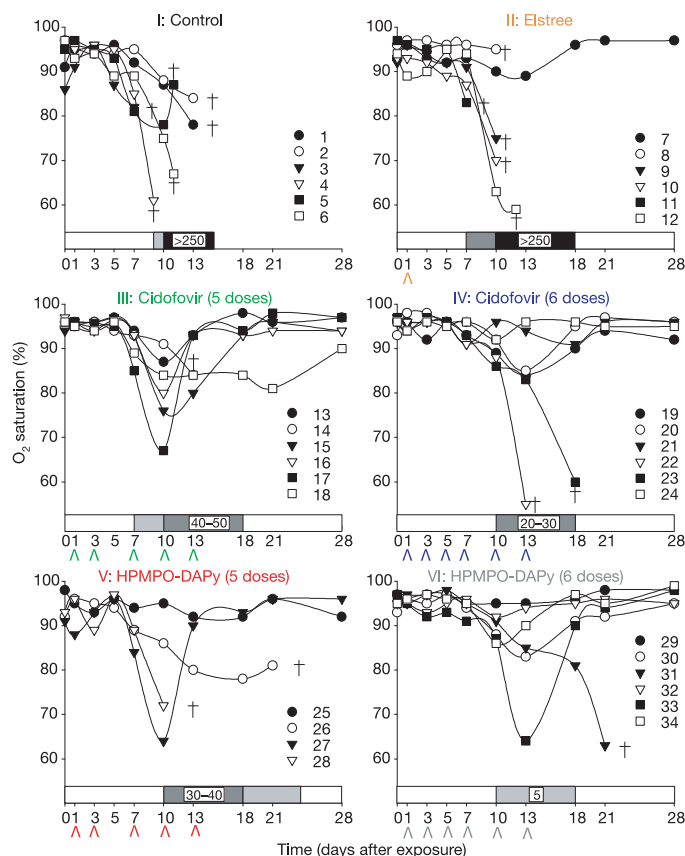


Figure 2 | Blood O_2 saturation profiles. Monkeys treated as described in Fig. 1. Individual monkeys (1–34) represented by symbols as indicated. Horizontal bar shows percentage of animals with cutaneous monkeypox lesions: none (white), 16–33% (light grey), 33–66% (dark grey), 66–100% (black). The cumulative number of cutaneous monkeypox lesions in all animals per group is shown in the bars. Crosses indicate time of death of individual animals. (See Supplementary Information for more details.)

(six-dose regimen) from severe morbidity and death (Fig. 1). Despite cidofovir treatment, mild morbidity was observed in five animals in group III and two animals in group IV (Fig. 2 and Supplementary Information). Similarly, treatment with HPMPO-DAPy protected two animals in group V (five-dose regimen, group size $n = 4$) and five animals in group VI (six-dose regimen, group size $n = 6$) from severe morbidity and death (Figs 1, 2 and Supplementary Information). Monitoring of plasma creatinine levels indicated that nephrotoxicity did not contribute to the mortality rate of the monkeys treated with the antiviral compounds (Supplementary Information).

Plasma MPXV loads were measured in all of the animals throughout the course of the experiment, using real-time polymerase chain reaction (PCR) (Fig. 3a). Viral loads largely paralleled the severity of disease and clinical signs. Five days after infection with MPXV, no significant differences in plasma viral load were found between control (group I) and vaccinated animals (group II, $P = 0.982$; Fig. 3b). Animals treated with cidofovir (groups III and IV) had significantly lower plasma viral loads than control and the vaccinated animals ($P = 0.021$ and 0.002 , respectively). Plasma viral loads in animals treated with HPMPO-DAPy (groups V and VI) were significantly lower than those of all the other groups ($P < 0.005$ – 0.041). Seven days after infection, plasma viral loads in animals treated six times with either cidofovir (group IV) or with HPMPO-DAPy (group VI) were significantly lower than those in the control or vaccinated animals ($P < 0.005$ and $P = 0.007$, respectively). The additional dose of cidofovir and HPMPO-DAPy given to groups IV and VI compared with groups III and V resulted in a significant

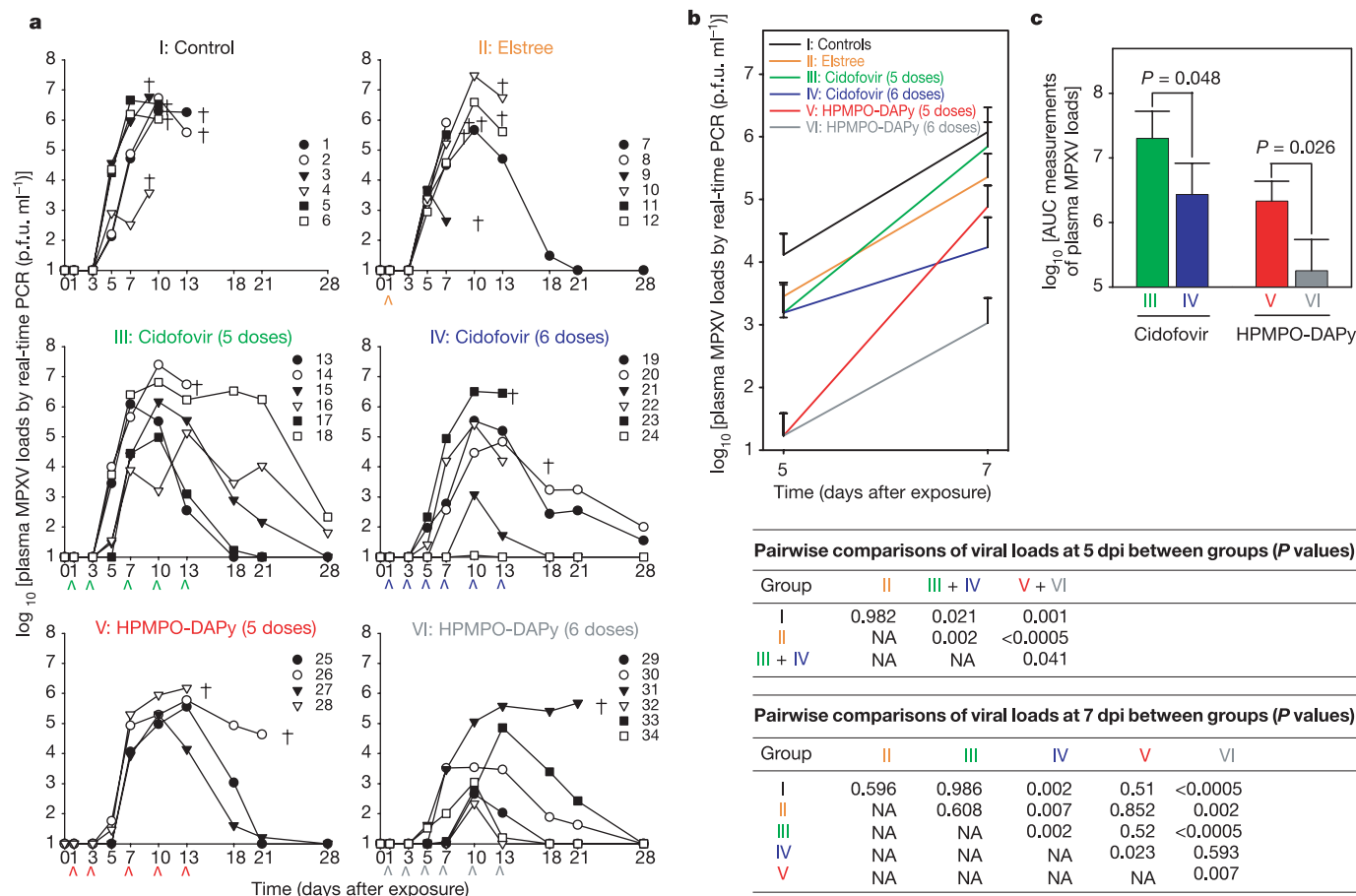


Figure 3 | Viral load profiles after infection with MPXV. a, Plasma MPXV loads of monkeys treated as described in Figs 1, 2. **b**, Plasma MPXV loads of monkeys (treated as described in Fig. 1) 5 or 7 days after infection. Data show average $\pm$ s.e.m. Statistical analysis of the pairwise comparisons of survival

additional reduction in plasma viral loads (cidofovir, $P = 0.002$; HPMPO-DAPy, $P = 0.007$). This was further corroborated by comparing the area under the curves for the respective groups until day 13 after infection, when all animals from groups III–VI were still alive (Fig. 3c; cidofovir, $P = 0.048$; HPMPO-DAPy, $P = 0.026$).

Antiviral treatment was stopped 13 days after infection, at which point all surviving animals had MPXV-specific plasma IgG antibodies, as measured by enzyme-linked immunosorbent assay (ELISA), as well as MPXV-specific T cells in their peripheral blood mononuclear cells, as measured by interferon- γ ELISPOT assay (Supplementary Information).

Taken together, our data show that post-exposure antiviral treatment with cidofovir or the related analogue HPMPO-DAPy, initiated 24 h after lethal respiratory infection of monkeys with MPXV, is more effective at reducing the mortality and number of cutaneous monkeypox lesions than standard vaccination with a classical variola vaccine administered 24 h after infection. Although mortality is a relevant read-out parameter, it should be noted that this is influenced largely by factors related to individual variation in pathogenesis and pathology. Given the relatively low numbers of animals that can be tested in non-human primate experiments, we also used other read-out criteria, including clinical, virological and immunological parameters. Of these, the number of cutaneous monkeypox lesions observed and the viral load data in the respective groups are probably the most relevant.

As the experimental compound HPMPO-DAPy was at least as effective as cidofovir in reducing numbers of cutaneous monkeypox lesions and viral loads, the potential usefulness of this compound

rates between groups is also shown. NA, not applicable. **c**, Area under the curve (AUC) measurements of plasma MPXV loads between days 0 and 13 after infection of monkeys treated as described in Fig. 1. Data show average $\pm$ s.e.m.

warrants further study. In this study, the monkeys received cidofovir or HPMPO-DAPy doses equivalent to those of cidofovir treatments given to humans, as well as the necessary supportive treatment with probenecid and extra hydration. We are therefore confident that the experimental conditions closely mimicked those for treatment of man after poxvirus infection²⁹.

In conclusion, this proof-of-concept study indicates that, in an outbreak situation, the use of cidofovir or a related nucleotide analogue with an improved bioavailability and toxicity profile, should be seriously considered an alternative or additional strategy for vaccination of individuals who may have been exposed to smallpox or other poxvirus infections such as monkeypox. Our study emphasizes the potential of cidofovir in the short-term treatment of acute poxvirus infections. For use against such infections in man, the treatment schedules used in this experimental model will require further optimization. Clinical experience with cidofovir is mainly based on long-term treatment of herpesvirus infections.

METHODS

Animals, MPXV infection, vaccination, antiviral treatment and sampling.

The studies were carried out in 34 captive-bred subadult (2–4 yr) healthy cynomolgus macaques (*Macaca fascicularis*) in accordance with the institutional guidelines for care and use of laboratory animals, which comply with legal requirements in The Netherlands. MPXV strain MSF#6, which was obtained from a fatally infected human in Congo, was provided by H. Meyer³⁰. The virus was passaged in African green monkey kidney cells: twice in MA104 cells and twice in Vero cells. The challenge virus stock had a titre of 2×10^8 p.f.u. per ml in Vero cell monolayers. Monkeys were inoculated intratracheally with 10^7 p.f.u. of the virus in a volume of 5 ml PBS. Animals died naturally (animals 2, 3, 8, 9, 10,

11, 18, 22, 23, 26, 28 and 31) or were killed for ethical reasons when they were moribund (animals 1, 4, 5, 6 and 12).

The classical smallpox vaccine (Lot. SP003, production date July 2002), referred to as Elstree-RIVM, was produced by the The National Institute for Public Health and the Environment (Bilthoven, The Netherlands). Elstree-RIVM was reconstituted as recommended by the manufacturer to give a final concentration of 10^8 TCID₅₀ ml⁻¹ (TCID₅₀, 50% tissue-culture infectious dose). A bifurcated needle was used to deliver the human equivalent dose of 2.5×10^5 TCID₅₀ intracutaneously. Cidofovir (provided by Gilead Sciences) and HPMPO-DAPy [(R)-2,4-diamino-6-[3-hydroxy-2-(phosphonomethoxy)propoxy]pyrimidine were weighed and dissolved in Dulbecco's modified Eagle's medium (Cambrex Bio Science) to yield a dose of 5 mg per kg body weight that was administered i.p. (after adjusting the pH with NaOH) in a volume of 2 ml per kg body weight on days 1, 3, 7, 10 and 13 after infection (five-dose regimen) or on days 1, 3, 5, 7, 10 and 13 after infection (six-dose regimen).

Additionally, 30 ml Ringer's solution was administered i.p. on each of these days to the animals treated with antivirals. These animals were also given oral probenecid (25 mg per kg body weight; Duchefa Biochemie B.V.) on days 0, 1, 3, 5 and 7 after infection. Probenecid was not administered on days 10 and 13 because of the risk of aspiration pneumonia in animals that showed severe dyspnea. Throat swabs and EDTA-anticoagulated blood samples were collected on days -1, 0, 1, 3, 5, 7, 10, 13, 18, 21 and 28 after infection.

Blood O₂ saturation. Blood (haemoglobin) O₂ saturation was measured using a pulse oximeter (Novamatrix Medical Systems Inc) at the times of sample collection, and expressed as the percentage of haemoglobin carrying O₂.

Virus quantification by real-time PCR. MPXV DNA was quantified in plasma and throat swabs using a LightCycler (Roche) as previously described².

Statistical analysis. Data were analysed using SPSS v.11. The outcome variables of viral load (at 5 and 7 days after infection) and the area under the curve of viral load (from 0 to 13 days after infection) were analysed after logarithmic (base 10) transformation. We assumed that after log transformation, viral loads had a normal distribution in the separate groups as well as in the combined groups. This assumption was verified by inspection of the distribution (mainly concerning symmetry) and using the Kolmogorov-Smirnov test, yielding *P*-values of about 0.70. We therefore used a one-way analysis of variance (ANOVA) to analyse the data.

There were six groups of monkeys in total. For viral load 5 days after infection, four groups were distinguished: groups III and IV were combined, and groups V and VI were combined because until day 5 after infection these animals received exactly the same treatments. Four groups were also considered for area under the curve measurements: groups I and II were excluded from this analysis because most of the animals had died by day 13 after infection. In the one-way ANOVA we first tested homogeneity of the within-group variances across the groups. Then, depending on the appropriate assumption of either homogeneous or heterogeneous variances, differences in means between the groups were tested. If and only if the overall group effect (that is, comparing all groups simultaneously) turned out to be significant (*P* < 0.05), a number of relevant pairwise comparisons were tested. Differences in survival between the groups were tested using the log-rank test. Again, pairwise comparisons were only tested if the overall *P* value from comparing all groups was significant (*P* < 0.05).

Received 10 September; accepted 29 September 2005.

Published online 11 December 2005.

1. Zaucha, G. M., Jahrling, P. B., Geisbert, T. W., Swearingen, J. R. & Hensley, L. The pathology of experimental aerosolized monkeypox virus infection in cynomolgus monkeys (*Macaca fascicularis*). *Lab. Invest.* **81**, 1581–1600 (2001).
2. Stittelaar, K. J. et al. Modified vaccinia virus Ankara protects macaques against respiratory challenge with monkeypox virus. *J. Virol.* **79**, 7845–7851 (2005).
3. Smith, G. L. & McFadden, G. Smallpox: anything to declare? *Nature Rev. Immunol.* **2**, 521–527 (2002).
4. Breman, J. G. & Arita, I. The confirmation and maintenance of smallpox eradication. *N. Engl. J. Med.* **303**, 1263–1273 (1980).
5. Lewis-Jones, S. Zoonotic poxvirus infections in humans. *Curr. Opin. Infect. Dis.* **17**, 81–89 (2004).
6. Grabenstein, J. D. & Winkenwerder, W. Jr US military smallpox vaccination program experience. *J. Am. Med. Assoc.* **289**, 3278–3282 (2003).
7. De Clercq, E. & Neyts, J. Therapeutic potential of nucleoside/nucleotide analogues against poxvirus infections. *Rev. Med. Virol.* **14**, 289–300 (2004).

8. Finkel, E. Australia. Engineered mouse virus spurs bioweapon fears. *Science* **291**, 585 (2001).
9. Stanford, M. M. & McFadden, G. The 'supervirus'? Lessons from IL-4-expressing poxviruses. *Trends Immunol.* **26**, 339–345 (2005).
10. Bray, M. & Roy, C. J. Antiviral prophylaxis of smallpox. *J. Antimicrob. Chemother.* **54**, 1–5 (2004).
11. Buller, R. M. et al. Efficacy of oral active ether lipid analogs of cidofovir in a lethal mousepox model. *Virology* **318**, 474–481 (2004).
12. Quenelle, D. C., Collins, D. J. & Kern, E. R. Efficacy of multiple- or single-dose cidofovir against vaccinia and cowpox virus infections in mice. *Antimicrob. Agents Chemother.* **47**, 3275–3280 (2003).
13. Roy, C. J., Baker, R., Washburn, K. & Bray, M. Aerosolized cidofovir is retained in the respiratory tract and protects mice against intranasal cowpox virus challenge. *Antimicrob. Agents Chemother.* **47**, 2933–2937 (2003).
14. Smee, D. F., Bailey, K. W. & Sidwell, R. W. Comparative effects of cidofovir and cyclic HPMPD on lethal cowpox and vaccinia virus respiratory infections in mice. *Chemotherapy* **49**, 126–131 (2003).
15. Neyts, J., Leyssen, P., Verbeken, E. & De Clercq, E. Efficacy of cidofovir in a murine model of disseminated progressive vaccinia. *Antimicrob. Agents Chemother.* **48**, 2267–2273 (2004).
16. Smee, D. F., Bailey, K. W., Wong, M. H., Wandersee, M. K. & Sidwell, R. W. Topical cidofovir is more effective than is parenteral therapy for treatment of progressive vaccinia in immunocompromised mice. *J. Infect. Dis.* **190**, 1132–1139 (2004).
17. De Clercq, E. Cidofovir in the therapy and short-term prophylaxis of poxvirus infections. *Trends Pharmacol. Sci.* **23**, 456–458 (2002).
18. Bronze, M. S. & Greenfield, R. A. Therapeutic options for diseases due to potential viral agents of bioterrorism. *Curr. Opin. Investig. Drugs* **4**, 172–178 (2003).
19. Prichard, M. N. & Kern, E. R. Orthopoxvirus targets for the development of antiviral therapies. *Curr. Drug Targets Infect. Disord.* **5**, 17–28 (2005).
20. Wharton, M. et al. Recommendations for using smallpox vaccine in a pre-event vaccination program. Supplemental recommendations of the Advisory Committee on Immunization Practices (ACIP) and the Healthcare Infection Control Practices Advisory Committee (HICPAC). *MMWR Recomm. Rep.* **52**, 1–16 (2003).
21. Yang, G. et al. An orally bioavailable anti-poxvirus compound (ST-246) that inhibits extracellular virus formation and protects mice from lethal orthopoxvirus challenge. *J. Virol.* **79**, 13139–13149 (2005).
22. Reeves, P. M. et al. Disabling poxvirus pathogenesis by inhibition of Abl-family tyrosine kinases. *Nature Med.* **11**, 731–739 (2005).
23. Yang, H. et al. Antiviral chemotherapy facilitates control of poxvirus infections through inhibition of cellular signal transduction. *J. Clin. Invest.* **115**, 379–387 (2005).
24. McFadden, G. Gleevec casts a pox on poxviruses. *Nature Med.* **11**, 711–712 (2005).
25. Balzarini, J. et al. 6-[2-phosphonomethoxy]alkoxy]-2,4-diaminopyrimidines: a new class of acyclic pyrimidine nucleoside phosphonates with antiviral activity. *Nucleosides Nucleotides Nucleic Acids* **23**, 1321–1327 (2004).
26. Neyts, J., Leyssen, P., De Clercq, E. & Holy, A. Effects of (S)-HPMPDAP and (S)-HPMPDAPy on vaccinia virus infections in mice. *Antiviral Res.* **62**, 109 (2004).
27. Rosenthal, S. R. & Clifford, J. C. Development of vaccines for bio-warfare agents. *Dev. Biol. (Basel)* **110**, 99–105 (2002).
28. Massoudi, M. S., Barker, L. & Schwartz, B. Effectiveness of postexposure vaccination for the prevention of smallpox: results of a delphi analysis. *J. Infect. Dis.* **188**, 973–976 (2003).
29. Cundy, K. C. Clinical pharmacokinetics of the antiviral nucleotide analogues cidofovir and adefovir. *Clin. Pharmacokinet.* **36**, 127–143 (1999).
30. Meyer, H. et al. Outbreaks of disease suspected of being due to human monkeypox virus infection in the Democratic Republic of Congo in 2001. *J. Clin. Microbiol.* **40**, 2919–2921 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We are grateful to F. Pistor, J. Rimmelzwaan-Offutt, S. Slabbekeorn-Romijn, R. D. d'Ullois and J. Guldemeester for assistance and to R. Gruters for comments. This work in Leuven and Prague was supported by an NIAID/NIH grant to J.N. and A.H.. This study is a part of an IOCB (Prague) research programme and project.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to A.D.M.E.O. (a.osterhaus@erasmusmc.nl).

Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat

Simon Griffiths¹, Rebecca Sharp¹, Tracie N. Foote¹, Isabelle Bertin¹, Michael Wanous², Steve Reader¹, Isabelle Colas¹ & Graham Moore¹

The foundation of western civilization owes much to the high fertility of bread wheat, which results from the stability of its polyploid genome. Despite possessing multiple sets of related chromosomes, hexaploid (bread) and tetraploid (pasta) wheat both behave as diploids at meiosis. Correct pairing of homologous chromosomes is controlled by the *Ph1* locus¹. In wheat hybrids, *Ph1* prevents pairing between related chromosomes². Lack of *Ph1* activity in diploid relatives of wheat suggests that *Ph1* arose on polyploidization³. Absence of phenotypic variation, apart from dosage effects, and the failure of ethylmethane sulphonate treatment to yield mutants, indicates that *Ph1* has a complex structure^{4,5}. Here we have localized *Ph1* to a 2.5-megabase interstitial region of wheat chromosome 5B containing a structure consisting of a segment of subtelomeric heterochromatin that inserted into a cluster of *cdc2*-related genes after polyploidization. The correlation of the presence of this structure with *Ph1* activity in related species, and the involvement of heterochromatin with *Ph1* (ref. 6) and *cdc2* genes with meiosis, makes the structure a good candidate for the *Ph1* locus.

The *Ph1* locus lies on chromosome 5B of hexaploid (AABBDD) and tetraploid (AABB) wheats, originally delimited by two overlapping deletions (*ph1b* and *ph1c*, respectively)^{7,8}. Gene order on chromosomes 5A, 5D and on the chromosomes of diploid relatives of polyploid wheat is very similar to that on chromosome 5B, but *Ph1* activity is unique to 5B. This suggests that *Ph1* arose through a structural change on 5B after polyploidization. Evidence indicating that this change was complex includes the failure of ethylmethane sulphonate (EMS) treatment to yield *Ph1* mutants⁴, and the fact that supernumerary chromosomes (heterochromatin) can compensate for the absence of *Ph1* (ref. 6). Moreover, duplication of the whole locus gives a stronger *Ph1* phenotype⁹, which would have been preferentially selected during wheat evolution if *Ph1* was a single gene effect. *Ph1* has pronounced effects both on the condensation of heterochromatin in hybrids of wheat with other Triticeae species, and on the condensation of euchromatin in wheat^{9,10}. Thus, the *Ph1* locus is likely to comprise a multigene family, heterochromatin or both, and affect condensation.

Mapping *Ph1* has been complicated by the absence of allelic variation, making it only possible to score for its presence or absence. This led us to propose a two-part strategy for the molecular characterization of the *Ph1* locus^{11–13}. The first part revealed the gene content of the wheat *Ph1* region using the conservation of gene order between the highly repetitive, 16,000-megabase (Mb) hexaploid wheat genome and the smaller genomes of rice and *Brachypodium sylvaticum* (460 Mb and 470 Mb, respectively), thus providing markers with which to saturate the wheat region. The second part of the strategy used deletion lines to dissect physically the *Ph1* locus. The original (*ph1b*) deletion in hexaploid wheat is some 70 Mb in

size, with the equivalent region in rice being approximately 4 Mb of chromosome 9 (ref. 14). Previously, fast-neutron irradiation of hexaploid wheat produced five deletions overlapping the *ph1b* deletion¹⁵. Using markers derived from the sequence of rice chromosome 9, we show that the breakpoints for these five wheat *Ph1* deletions cluster within a 250-kilobase (kb) section of the 4-Mb equivalent region of rice chromosome 9. These breakpoints delimit *Ph1* to a location equivalent to a 140-kb section of the rice chromosome (Supplementary Fig. 1). The nonrandom clustering of the breakpoints suggests that the whole segment is in a different conformation to its flanking regions.

All markers derived from the 140-kb rice region were mapped against the wheat *Ph1* deletion lines to confirm their location. Half of the rice markers failed to yield clear hybridizing patterns on wheat Southern blots. Therefore, the rice markers were used to screen a *B. sylvaticum* bacterial artificial chromosome (BAC) library to identify homologous markers from a closer relative of wheat¹⁶. Screening this library also identified additional markers with which to analyse the wheat region. The rice markers identified two regions in *Brachypodium*: region 1 is co-linear with the *Ph1* region on rice chromosome 9, and region 2 with rice chromosome 8 (Fig. 1 and Supplementary Fig. 2). This is consistent with known internal duplications in the rice genome¹⁷. In contrast to the rice markers, all the *Brachypodium* markers gave clear hybridizing patterns when mapped against the deletion lines.

Forty-two markers, mostly from *Brachypodium* region 1 genes, were used to screen hexaploid and tetraploid wheat BAC libraries containing 750,000 and 500,000 clones, respectively^{18,19}. Southern blots, prepared from restriction digests of 2,000 positive BACs and a full set of wheat nullisomic/tetrasomic lines and deletion lines, were hybridized with the same probes. The hexaploid wheat BAC library is made from the same variety of wheat as the nullisomic/tetrasomic lines, so BACs could be unambiguously assigned to wheat chromosomes by comparison of hybridizing fragment size. This process eliminated gene family members from other genomic locations, giving 400 BACs from the *Ph1* region and the homologous regions of chromosomes 5A and 5D. Overlaps were confirmed by cross-hybridization of end clones and comparison of restriction profiles. A minimum tiling path from these contigs was selected for shotgun sequencing. Thirty-six genes were identified, 25 of which were co-linear with their orthologues on rice chromosome 9 and/or *Brachypodium* region 1. The generation of markers from this sequence enabled some gaps in the contigs to be closed through chromosome walking. After all possibilities for further steps were stopped by large tracts of retroelements, we had assembled five BAC contigs for the region, with approximate sizes of 500 kb, 300 kb, 1.2 Mb, 1 Mb and 200 kb (Fig. 1). More sequencing was undertaken to reveal the gene content of the areas we had walked into, making a

¹John Innes Centre, Colney, Norwich NR4 7UH, UK. ²Augustana College, Sioux Falls, South Dakota 57197, USA.

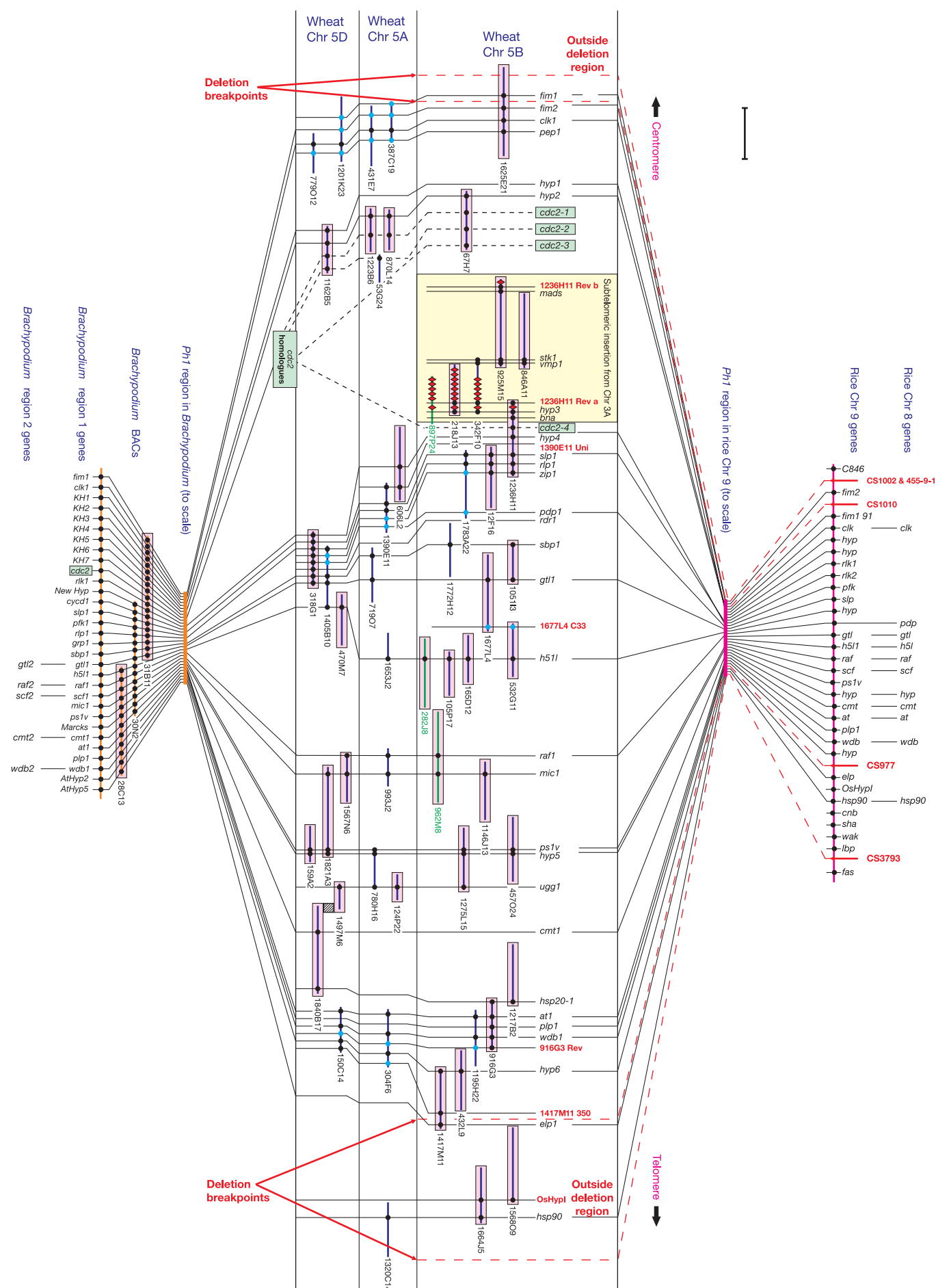


Figure 1 | BAC tiling path and annotated gene content of the *Ph1* region and its equivalent region in rice chromosomes 9 and 8 and *Brachypodium* regions 1 and 2. Pink vertical lines represent part of rice chromosome 9; orange vertical lines represent *Brachypodium* chromosome region 1. Co-linearity of rice and *Brachypodium* markers with markers on part of wheat chromosomes 5B, 5A and 5D is shown with horizontal and diagonal black lines; red dashed lines show the location of fast neutron-induced breakpoints on chromosome 5B. Blue bars represent hexaploid wheat BACs; green bars show tetraploid wheat BACs (*T. durum*). Pink boxes indicate sequenced BACs. Gene and BAC names are indicated in black, whereas other

total of 33 *Triticum aestivum* BACs sequenced within this region. So far, this represents the largest study of its kind in wheat.

All the genes within the *Ph1* 5B deletion region (as defined by deletion line breakpoints) were assessed for their expression in different tissues (leaf, spike and seed) and, where expression was detected, none of them showed differences in the presence and absence of *Ph1* (Supplementary Table 1 and Supplementary Fig. 3). However, on the basis of their sequence homology to genes of known function, the *cdc2* genes are the best candidates for *Ph1* function. *cdc2*-related genes affect chromosome condensation (<http://www.ihop-net.org/UniPub/iHOP/>). This is the only multigene cluster in the region, and at least one of its members (*cdc2-4*) is 5B specific. It is striking that this cluster of *cdc2* genes is interrupted by a large segment of chromosome 5B (yellow boxed area in Fig. 1) that does not have homologous equivalents on 5A and 5D. Southern analysis shows that the genes in this additional segment are all part of a subtelomeric insertion from chromosome 3AL. It also contains two tandem arrays of a 2.3-kb repeating unit, showing no sequence homology to previously recorded repetitive elements. Our analysis indicates that the tandem array between *vmp1* and *hyp3* genes is at least 87 kb in size, and between *cdc2-3* and the *mads* pseudogene is at least 20 kb. Some of the BACs carrying this repeat were unstable. Southern analysis reveals that the tandem arrays are not present on wheat chromosomes 5A or 5D, but there are a number of similar repeats on chromosome 3B, also occurring as arrays (Supplementary Fig. 4). These data indicate that an insertion from the subtelomeric region of 3AL has occurred in this part of the newly defined *Ph1* region. The novel tandem repeat together with *cdc2* genes fulfils all our criteria as a *Ph1* candidate structure.

In situ hybridization of the tandem repeat on mitotic chromosome spreads from hexaploid wheat root meristems reveals two strongly hybridizing interstitial and two weaker hybridizing subtelomeric locations (Supplementary Fig. 5). The two interstitial sites correspond to the tandem arrays within the *Ph1* locus on chromosome 5B, whereas the subtelomeric sites on the long arms represent the 3B arrays. In the absence of *Ph1*, only the subtelomeric sites can be visualized in meiocytes. The availability of markers for the 3BL telomeric regions enables us to show that these regions pair in all meiocytes in the presence and absence of *Ph1* in hexaploid wheat (Supplementary Fig. 5a–f). Thus, telomeres pair correctly independently of *Ph1* activity.

If the segment containing the tandem repeats and the *cdc2-4* (5B specific) pseudogene is responsible for the *Ph1* effect, it should be present in relatives of hexaploid wheat possessing *Ph1*. Chromosome 5B of tetraploid wheat also possesses a *Ph1* locus defined by the *ph1c* deletion⁸. Southern analysis shows that the additional segment is also within the *ph1c* deletion (Supplementary Fig. 4). There is a high level of conservation between the restriction profiles of tetraploid wheat BACs carrying the tandem repeat and those derived from the corresponding hexaploid wheat 5B region (Supplementary Fig. 6). Although the S genome of *Aegilops speltoides* (SS) is the progenitor of the B genome of *Triticum dicoccoides* (AABB) and the G genomes of *Triticum timopheevi* (AAGG) and *Triticum araraticum* (AAGG), it does not possess *Ph1* activity, but the B and G genomes of these tetraploids do. Moreover, *T. araraticum* and *T. timopheevi* result from

marker names are in red. Black circles indicate the presence of marker from sequence data or by Southern analysis; blue circles show presence of marker by colony hybridization alone; no circle indicates not tested. The yellow box represents a subtelomeric insertion from wheat chromosome 3A. Red diamonds represent tandem repeats. The high level of homology between *cdc2*-type genes on different BACs means that homologues cannot be distinguished from paralogues; they are therefore connected by black dashed lines. The hashed box indicates overlap between BACs based on fingerprint only. Centromere and telomere directions are indicated with arrows. Scale bar, 100 kb. All wheat BACs are drawn to scale.

a later polyploidization event than their B genome relatives. Despite their separate polyploidization history, polymerase chain reaction (PCR) assays that are 5B specific for the *cdc2-4* pseudogene and 2.3-kb repeat (data not shown) in *T. aestivum* amplified products from all of the *T. dicoccoides*, *T. timopheevi* and *T. araraticum* accessions tested, but failed to amplify products from *A. speltoides* and its relative *Aegilops longissima* (S¹S¹) (Supplementary Fig. 7). Lack of the *cdc2-4* pseudogene in the B genome progenitor (the S genome) indicates that it is part of the polyploidization event. The availability of these markers will enable us to understand how *Ph1* arose in these species. Fertile hybrids have been generated between accessions of *T. timopheevi* and *T. dicoccoides*, so genetic exchange is possible²⁰.

Tight coupling of the *cdc2*/heterochromatin structure with *Ph1* activity leads to the conclusion that they are essential components of this locus. The presence of a multigene complex containing heterochromatin, and the potential lethality of *cdc2* mutants, means that the dissection of the *Ph1* locus by a transgenic approach is difficult. However, the use of drugs such as okadaic acid, which alter *cdc2* activity, and a multiplex PCR assay of markers across the locus (Supplementary Fig. 8) to identify deletions of the segment, will allow dissection of *Ph1* function and the generation of *Ph1* variation essential for breeding.

METHODS

Plant material. The following hexaploid wheat (*Triticum aestivum*; $2n = 6 \times = 42$; genome AABBDD) lines were used: euploid Chinese Spring, Chinese Spring mutant line *ph1b*, Chinese Spring nullisomic/tetrasomic lines N5AT5B, N5BT5D and N5DT5A, Chinese Spring 3BL deletion line 4525L7 (ref. 21), Highbury (CS 5B) *Ph1* mutant lines 977, 1002, 1010, 3793 and 455-9-1. These genetic stocks have been previously described¹⁵. Tetraploid wheat lines used were: *Triticum durum* ($2n = 4 \times = 28$ AABB), Cappelli and its derivative deletion line *ph1c*, *T. timopheevi* ($2n = 4 \times = 28$ AAGG), *T. araraticum* ($2n = 4 \times = 28$ AAGG) and *T. dicoccoides* ($2n = 4 \times = 28$ AABB). Specific *T. araraticum* and *T. dicoccoides* accessions used are shown in Supplementary Fig. 7. Diploid wheat species used were: *Triticum monococcum* ($2n = 2 \times = 14$ AA), *A. longissima* ($2n = 2 \times = 14$ S¹S¹), *A. speltoides* ($2n = 2 \times = 14$ SS) and *Triticum tauschii* ($2n = 2 \times = 14$ DD).

BAC libraries and their analysis. The hybridization and analysis of the wheat and *Brachypodium* BAC libraries have been previously described^{16,18}.

Sequencing and annotation. Sequencing was performed by GATC Biotech Ltd using standard techniques. Contigs were assembled using PhredPhrap^{22,23} and viewed using Consed²⁴. Gene models were based on *ab initio* Genemark²⁵ predictions, sequence conservation between wheat, rice and *Brachypodium sylvaticum*, and homology to known peptides detected by BLAST²⁶. Intron-exon boundary predictions were refined using Geneserper²⁷. Sequences were submitted to EMBL/GenBank/DDJB in HTG format with an average of 35 contigs per BAC. Accession numbers, gene content and sequence coverage are shown in Supplementary Table 2.

Fluorescence *in situ* hybridization. Root tips and anthers were collected as previously described^{9,10}. Meiocytes were extruded from individual anthers and spread on slides²⁸. Primers were designed from the 5B *Ph1* region tandem repeat and a 2-kb product was amplified by PCR using BAC 218J13 DNA as template. Probes were prepared and *in situ* hybridization was carried out as previously described^{9,10}.

Fluorescence microscopy and image processing. Cells were visualized using a Nikon Eclipse E600 fluorescence microscope connected to a Hamamatsu CCD camera, and digital images were captured using MetaMorph (Universal Imaging

Corp.) software. Images were processed using the AutoDeblur (AutoQuant Imaging) and ImageJ (NIH) programs.

Analysis of expression by RT-PCR. Total RNA isolation, reverse transcription and measurement of the abundance of products using a Typhoon fluorimeter have all been previously described²⁹. Tissues included wheat leaves as described, immature inflorescences taken at Feekes scale 6–7 and seed imbibed for 24 h (Chinese Spring and *ph1b* types). PCR was performed on the cDNA using intron-spanning, gene-specific primer pairs. PCR products were cloned using a p-GEM T easy vector kit (Promega) according to the manufacturer's instructions.

Received 29 September; accepted 11 November 2005.

- Riley, R. & Chapman, V. Genetic control of cytologically diploid behaviour of hexaploid wheat. *Nature* **182**, 713–715 (1958).
- Riley, R., Chapman, V. & Kimber, G. Genetic control of chromosome pairing in intergeneric hybrids with wheat. *Nature* **183**, 1244–1246 (1959).
- Chapman, V. & Riley, R. Homoeologous meiotic chromosome pairing in *Triticum aestivum* in which chromosome 5B is replaced by an alien homoeologue. *Nature* **226**, 376–377 (1970).
- Wall, A. M., Riley, R. & Gale, M. D. The position of a locus on chromosome 5B of *Triticum aestivum* affecting homoeologous meiotic pairing. *Genet. Res. Camb.* **18**, 329–339 (1971).
- Martinez, M., Naranjo, T., Cuadrado, C. & Romero, C. The synapctic behaviour of *Triticum turgidum* with variable doses of the *Ph1* locus. *Theor. Appl. Genet.* **102**, 751–758 (2001).
- Dover, G. A. & Riley, R. Prevention of pairing of homoeologous meiotic chromosomes of wheat by an activity of supernumerary chromosomes of *Aegilops*. *Nature* **240**, 159–161 (1972).
- Sears, E. R. An induced mutant with homoeologous pairing in common wheat. *Can. J. Genet. Cytol.* **19**, 585–593 (1977).
- Giorgi, B. A homoeologous pairing mutant isolated in *Triticum durum* cv Cappelli. *Mutat. Breed. Newslett.* **11**, 4–5 (1978).
- Martinez-Perez, E., Shaw, P. & Moore, G. The *Ph1* locus is needed to ensure specific somatic and meiotic centromere association. *Nature* **411**, 204–207 (2001).
- Prieto, P., Shaw, P. & Moore, G. Homologue recognition during meiosis is associated with a change in chromatin conformation. *Nature Cell Biol.* **6**, 906–908 (2004).
- Moore, G., Gale, M. D., Kurata, N. & Flavell, R. B. Molecular analysis of small grain cereal genomes. *Nature Biotechnol.* **11**, 584–589 (1993).
- Kurata, N. *et al.* Conservation of genome structure between rice and wheat. *Nature Biotechnol.* **12**, 276–278 (1994).
- Sasaki, T., Matsumoto, T., Antonio, B. A. & Nagamura, Y. Review: from mapping to sequencing and beyond. *Plant Cell Physiol.* **46**, 3–12 (2005).
- Foot, T., Roberts, M., Kurata, N., Sasaki, T. & Moore, G. Detailed comparative mapping of cereal chromosome regions corresponding to the *Ph1* locus in wheat. *Genetics* **147**, 801–807 (1997).
- Roberts, M. A. *et al.* Induction and characterisation of *Ph1* wheat mutants. *Genetics* **153**, 1909–1918 (1999).
- Foot, T. N., Griffiths, S., Allouis, S. & Moore, G. Construction and analysis of a BAC library in the grass *Brachypodium sylvaticum*: its use as a tool to bridge the gap between rice and wheat in elucidating gene content. *Funct. Integr. Genomics* **4**, 26–33 (2004).
- Paterson, A. H., Bowers, J. E. & Chapman, B. A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci. USA* **101**, 9903–9908 (2004).
- Allouis, S. *et al.* Construction and characterisation of a hexaploid wheat (*Triticum aestivum* L.) BAC library from the reference germplasm 'Chinese Spring'. *Cereal Res. Commun.* **31**, 331–338 (2003).
- Cenci, A. *et al.* Construction and characterisation of a half million clone BAC library of durum wheat (*Triticum turgidum* ssp. *durum*). *Theor. Appl. Genet.* **107**, 931–939 (2003).
- Rao, P. S. & Smith, E. L. Studies with Israeli and Turkish accessions of *Triticum* Lemend. var. *dicoccoides* (KORN) BOWDEN. *Wheat Inform. Serv.* **26**, 6–7 (1968).
- Endo, T. R. & Gill, B. S. The deletion stocks of common wheat. *J. Hered.* **87**, 295–307 (1996).
- Ewing, B., Hillier, L., Wendl, M. & Green, P. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–185 (1998).
- Ewing, B. & Green, P. Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194 (1998).
- Gordon, D., Abajian, C. & Green, P. Consed: A graphical tool for sequence finishing. *Genome Res.* **8**, 195–202 (1998).
- Borodovsky, M. & McIninch, J. GeneMark: Parallel gene recognition for both DNA strands. *Comput. Chem.* **17**, 123–133 (1993).
- Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
- Schlueter, S. D., Dong, Q. F. & Brendel, V. GeneSequer@PlantGDB: gene structure prediction in plant genomes. *Nucleic Acids Res.* **31**, 3597–3600 (2003).
- Bass, H. W. *et al.* Evidence for the coincident initiation of homolog pairing and synapsis during the telomere-clustering (bouquet) stage of meiotic prophase. *J. Cell Sci.* **113**, 1033–1042 (2000).
- Dunford, R. P., Griffiths, S., Christodoulou, V. & Laurie, D. A. Characterisation of a barley (*Hordeum vulgare* L.) homologue of the *Arabidopsis* flowering time regulator *Gigantea*. *Theor. Appl. Genet.* **110**, 925–931 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was funded by the Biotechnology and Biological Sciences Research Council of the UK. Sabbatical support for M.W. was provided by a NIH grant from the NCRR INBRE program. We would like to thank DuPont/Pioneer Hi-Bred International and P. Prieto for their assistance.

Author Information Sequences were submitted to EMBL/GenBank/DDJB and accession numbers are shown in Supplementary Table 2. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to G.M. (graham.moore@bbsrc.ac.uk).

ClpS is an essential component of the N-end rule pathway in *Escherichia coli*

A. Erbs^{1*}, R. Schmidt^{1*}, T. Bornemann^{1†}, J. Schneider-Mergener², A. Mogk¹, R. Zahn¹, D. A. Dougan^{1,3} & B. Bukau¹

The N-end rule states that the half-life of a protein is determined by the nature of its amino-terminal residue¹. Eukaryotes and prokaryotes use N-terminal destabilizing residues as a signal to target proteins for degradation by the N-end rule pathway. In eukaryotes an E3 ligase, N-recogin, recognizes N-end rule substrates and mediates their ubiquitination and degradation by the proteasome^{1,2}. In *Escherichia coli*, N-end rule substrates are degraded by the AAA+ chaperone ClpA in complex with the ClpP peptidase (ClpAP)³. Little is known of the molecular mechanism by which N-end rule substrates are initially selected for proteolysis. Here we report that the ClpAP-specific adaptor, ClpS, is essential for degradation of N-end rule substrates by ClpAP in bacteria. ClpS binds directly to N-terminal destabilizing residues through its substrate-binding site distal to the ClpS–ClpA interface⁴, and targets these substrates to ClpAP for degradation. Degradation by the N-end rule pathway is more complex than anticipated and several other features are involved, including a net positive charge near the N terminus and an unstructured region between the N-terminal signal and the folded protein substrate. Through interaction with this signal, ClpS converts the ClpAP machine into a protease with exquisitely defined specificity, ideally suited to regulatory proteolysis.

In eukaryotes, the signal for N-end rule degradation (N-degron) comprises an N-terminal primary destabilizing amino acid (type 1, R, K, H; type 2, F, L, W, I, Y) and an accessible lysine for ubiquitination. N-terminal secondary and tertiary destabilizing residues initiate enzymatic attachment of primary destabilizing residues¹. N-recognins interact with destabilizing N termini through distinct binding sites for type-1 and type-2 destabilizing residues^{1,5} and target substrates for ubiquitination and proteasomal degradation. In prokaryotes, the degradation signal comprises N-terminal primary destabilizing residues (F, L, W, Y). Secondary destabilizing residues (R, K) recruit primary destabilizing residues (F, L) by conjugation by a F/L-transferase³. ClpA unfolds N-end rule substrates and translocates them to ClpP³. The recognition of these substrates by ClpA remains unclear. The ClpA adaptor protein ClpS binds to the N-domain of ClpA and inhibits ClpAP-mediated degradation of several substrates while stimulating degradation of model aggregates *in vitro*⁶. The ClpS-mediated degradation of aggregates is not significant *in vivo*⁶, suggesting that ClpS has an alternative function *in vivo*. Because it has been proposed that ClpS, despite low sequence homology, shares considerable secondary structure with the predicted secondary structure of the putative type-2 binding site of N-recognins⁷, we speculated that ClpS might target N-end rule substrates for degradation by ClpAP.

We tested the role of ClpS in N-end rule degradation *in vivo* by

using a previously developed genetic system³. We compared the stability of different X-β-galactosidase (X-β-gal, where X is the engineered N-terminal residue) constructs in wild-type and $\Delta clpS$ *E. coli* cells. As expected β-gal containing a stabilizing N-terminal methionine (M-β-gal) was stable in wild-type cells (Fig. 1a), whereas β-gal containing a primary (L-β-gal) or secondary (R-β-gal) destabilizing residue was rapidly degraded (Fig. 1b, c). Notably R-β-gal, processed to FR-β-gal or LR-β-gal by an F/L-transferase before proteolysis, was degraded more rapidly than L-β-gal. By contrast, all three β-gal proteins were stable in $\Delta clpS$ cells (Fig. 1a–c). The degradation of N-end rule β-gal proteins could be restored in $\Delta clpS$ cells by coexpression of ClpS (Fig. 1d), showing that ClpS is essential for ClpAP-dependent degradation of N-end rule proteins in *E. coli*.

To determine whether ClpS directly binds to the N-degron, we screened libraries of carboxy-terminally coupled peptides⁸. Peptides were derived from the ten N-terminal residues of green fluorescent protein (GFP; Fig. 2a) or β-gal (Supplementary Fig. S2a), and the

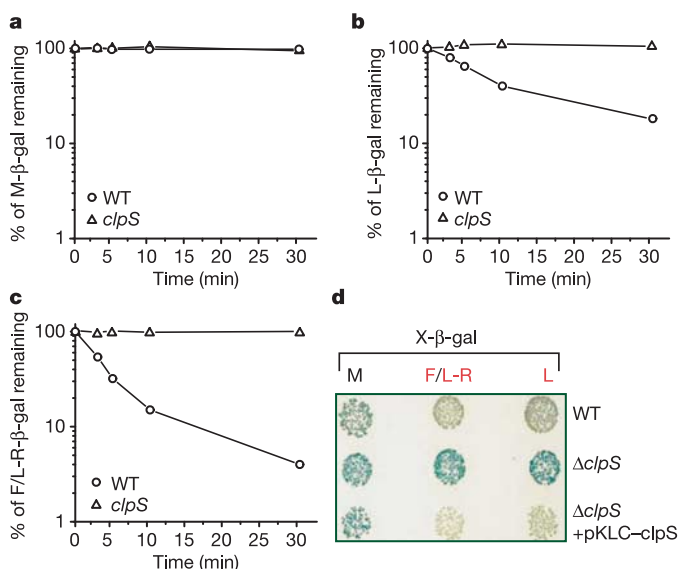


Figure 1 | ClpS is essential for ClpAP-mediated degradation of N-end rule substrates *in vivo*. **a–c**, Wild-type and $\Delta clpS$ mutant cells were pulse-labelled with [³⁵S]methionine and quantified for M-β-gal (**a**), L-β-gal (**b**) or F/L-R-β-gal (**c**). **d**, Expression of ClpS from plasmid pKLC-clpS complements the loss of degradation of N-end rule β-gal in $\Delta clpS$ mutant cells. Blue colonies indicate the presence of β-gal; white colonies indicate degradation of β-gal.

¹Zentrum für Molekulare Biologie Heidelberg, Universität Heidelberg, INF 282, Heidelberg D-69120, Germany. ²Universitätsklinikum Charité, Humboldt-Universität, Schumannstrasse 20–21 Berlin D-10098, Germany. ³Department of Biochemistry, La Trobe University, Melbourne 3086, Australia. [†]Present address: Institut für Molekularbiologie, Universität Witten/Herdecke, Stockumer Strasse 10–12, Witten D-58453, Germany.

*These authors contributed equally to this work.

N-terminal residues in the first and second position were systematically exchanged. ClpS bound strongly only to peptides with destabilizing primary residues (F, L, W, Y) at the N terminus (Fig. 2a and Supplementary Fig. S2a). The additional presence of a secondary destabilizing residue (R, K) strongly enhanced ClpS binding and even resulted in considerable binding if the primary destabilizing residue was replaced with R or K. The higher affinity of ClpS to peptides containing both primary and secondary destabilizing residues is consistent with the faster *in vivo* degradation of F/L-R- β -gal as compared with L- β -gal. This suggests that secondary destabilizing residues contribute directly to ClpS binding, in addition to initiating attachment of primary destabilizing residues. Notably, N-recognins, in contrast to ClpS, have evolved a separate binding site for positively charged N-terminal residues, allowing the recognition of L and R as primary destabilizing residues independent of the activity of a transferase¹.

To assess the role of the location of the destabilizing residues we moved the FR motif systematically towards the C terminus of the peptides (Fig. 2b and Supplementary Fig. S2b). Strong ClpS binding only occurred when the FR motif was located at the absolute N terminus of the peptide. Furthermore, acetylation of the α -amino group abolished ClpS binding (Fig. 2c), showing that the amino group of the N-terminal destabilizing residue is essential for recognition of the N-degron by ClpS.

To confirm the specificity of ClpS for N-end rule peptides we screened 508 peptides, scanning the sequences of three different proteins with an overlap of ten residues between neighbouring spots. Figure 2d shows the pattern of ClpS binding to a library from the λ CI DNA-binding domain. In contrast to other chaperones tested^{8–10}, ClpS bound primarily to isolated spots throughout the library corresponding to N-end rule peptides (Fig. 2d). Consistently, ClpS binding was abolished when the destabilizing amino acids were not located at the N terminus (Fig. 2d, spots 31, 32). Analysis of all ClpS-binding peptides showed that N-end rule peptides are highly enriched among the strongest ClpS binders (Fig. 2e), although some ClpS-binding peptides diverged from the N-end rule motif.

We found that strong N-end rule ClpS-binding peptides favour a positive net charge (Fig. 2f). Accordingly, a negative net charge or a clustering of negative charges towards the N terminus can inhibit ClpS binding (Fig. 2d, spots 29, 30, 48), although ClpS may still bind if the negative charges are located towards the C terminus or are shielded by neighbouring positive charges (Fig. 2f). Analysis of ClpS binders showed that a strong positive net charge may partially compensate for a missing primary destabilizing residue (for example, spots 6, 64 and 65 in Fig. 2d). Although peptides of this class are not recognized by ClpS as faithfully as N-degron peptides, the fact that some of them bind raises the possibility that ClpS can target other substrate classes ClpAP.

We investigated the mechanism of ClpS-dependent N-end rule degradation by ClpAP (ClpAPS) *in vitro* using a set of GFP variants (Fig. 3a). Wild-type GFP has stabilizing residues at the N terminus and is not degraded by ClpAP or ClpAPS. Initially, we fused the FR motif directly to the N terminus of GFP (Fig. 3a). FR-GFP was slowly degraded by ClpAP, and this degradation was completely inhibited by ClpS (Fig. 3b). We speculated that the N-terminal destabilizing residues may not be accessible for ClpS binding, converting ClpS into an inhibitor of ClpAP activity as it is for GFP-ssrA⁶. Similarly, it has been reported that a linker is needed to ensure accessibility of the N-degron to the N-recognin Ubr1 (ref. 11). We inserted a linker, comprising a repeat of the ten N-terminal residues of GFP, between the folded β -barrel and the N terminus of GFP either with (FR-Linker-GFP) or without (Linker-GFP) N-terminal destabilizing residues. ClpAP degraded Linker-GFP (Supplementary Fig. S3) and FR-Linker-GFP (Fig. 3c) with a slow rate similar to that found for degradation of FR-GFP. Addition of ClpS, however, drastically increased the degradation rate of FR-Linker-GFP (Fig. 3c), but blocked the minor degradation of Linker-GFP (Supplementary

Fig. S3). ClpS thus specifically mediates the degradation of N-end rule proteins by ClpAP and, at the same time, prevents the degradation of other substrates.

To investigate the role of the linker more closely, we determined whether ClpS interacts with FR-GFP or FR-Linker-GFP using surface plasmon resonance. Immobilized ClpS bound both FR-GFP and FR-Linker-GFP but not wild-type GFP (Fig. 3d). The dissociation constant (K_d) for FR-GFP (260 nM) was lower than that

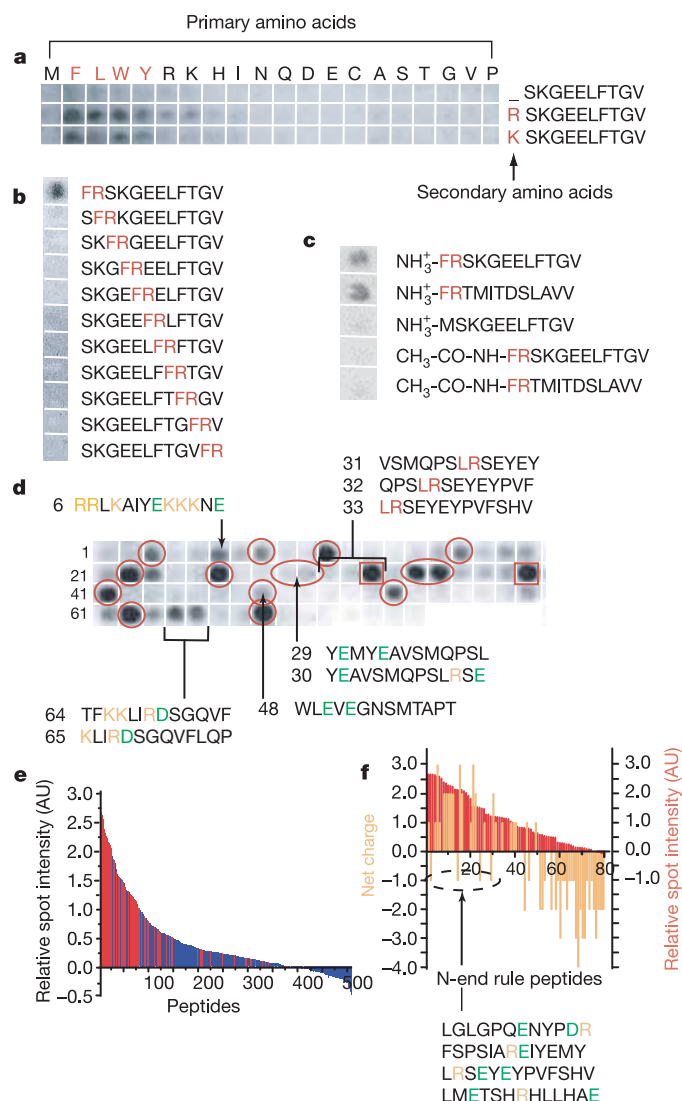


Figure 2 | ClpS binds to N-end rule peptides in peptide libraries.

Destabilizing residues are shown in red. **a**, ClpS binding to peptides derived from the N-terminal ten residues of GFP with added permuted primary and secondary residues. Letters at the top indicate the N-terminal residue; letters to the right indicate the peptide sequence in the marked row, starting with the second residue. **b**, ClpS binding to N-end rule peptides requires N-terminal positioning of destabilizing residues. Letters indicate the peptide sequences. **c**, ClpS does not bind to peptides with acetylated α -amino groups. **d**, Binding of ClpS to a library derived from the λ CI DNA-binding domain. Numbers to the left indicate the first peptide spot in the row. Red circles, peptides with primary destabilizing residues; red squares, peptides with primary and secondary destabilizing residues. Spots referred to in the text are given with their sequences. Red, destabilizing residues; green, negatively charged residues; yellow, positively charged residues. **e**, Binding affinities of ClpS to 508 peptides correlated to destabilizing N termini. Peptides are sorted by decreasing signal intensity (in arbitrary units). N-end rule binders (red bars) show higher affinity for ClpS than do non-N-end rule peptides (blue bars). **f**, Positive net charge of N-end rule peptides strengthens ClpS affinity. All N-end rule peptides tested were sorted by decreasing signal intensity (red bars) and plotted with their net charges (yellow bars).

for FR-Linker-GFP (1,450 nM; Supplementary Fig. S4a, b). We speculate that the change in K_d originates from an increase in accessibility of the negative charges located in the linker. These results establish that ClpS binds directly to both N-end rule substrates and exclude the possibility that the destabilizing terminus of FR-GFP is not accessible to ClpS. We propose that the linker is needed either to overcome steric restrictions during substrate transfer from ClpS to ClpA or because substrate transfer and/or unfolding requires simultaneous contact of the substrate to both ClpS and ClpA. It is important to note that the β -gal encoded in the ubiquitin-X- β -gal constructs used in the *in vivo* studies described here and elsewhere³ possesses a modified N-terminal sequence in which the N-terminal 19 residues of wild-type β -gal are replaced with a 46-residue sequence that can serve as a linker.

To gain structural insights into the molecular basis of substrate binding by ClpS, we compared the sequences of several ClpS and N-recognin homologues from different organisms (Fig. 4a) and identified a conserved region in ClpS, located on the surface of the protein, distal to the ClpA-ClpS interface. This region is complementary to the ClpS-binding motif that we have identified here and contains several surface exposed negatively charged residues (D35, D36 and E41) and a hydrophobic residue (Y37; Fig. 4b). The charged and hydrophobic character of this region is also conserved in the putative type-2 binding site of N-recognin⁷ (Fig. 4a). To investigate the role of these residues in N-end rule degradation, we constructed three ClpS mutants (Y37A, E41A and a double mutant D35A/D36A referred to hereafter as DD/AA) and tested their ability to mediate degradation of N-end rule substrate *in vitro*. Although all three ClpS mutants showed normal interaction with ClpA (Supplementary Fig. S5), their ability to mediate the degradation of N-end rule substrates was impaired (Fig. 4c). The single mutants (E41A and Y37A) were partially impaired in their ability to mediate degradation of FR-Linker-GFP *in vitro* (Fig. 4c), whereas the double-mutant DD/AA was completely defective (Fig. 4c). Consistently, expression of DD/AA in Δ clpS cells (Fig. 4d) did not restore the degradation of N-end rule substrates, L- β -gal or F/L-R- β -gal, whereas expression of ClpS-E41A or ClpS-Y37A showed only a weak phenotype. These results indicate that the highly conserved residues (D35 and D36) form a crucial part of the N-end rule substrate-binding site in ClpS. We speculate that the interaction of the α -amino group of the primary destabilizing residue with the negative charges in the ClpS binding pocket drives binding of the N terminus of N-end rule substrates and that the bulky, hydrophobic side chain is needed for the correct positioning of the N terminus.

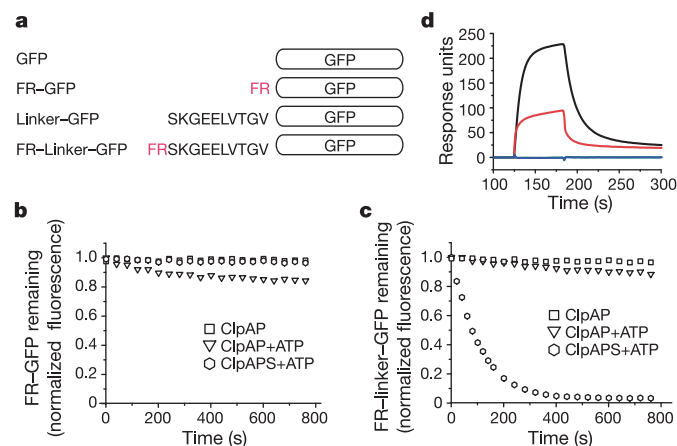


Figure 3 | ClpS mediates degradation of a N-end rule GFP variant by ClpAP *in vitro*. **a**, GFP constructs used to test N-end rule degradation by ClpAPS. **b**, **c**, Degradation of the FR-GFP (**b**) and FR-Linker-GFP (**c**) monitored by fluorescence. **d**, Biacore sensogram showing binding of FR-Linker-GFP (red) and FR-GFP (black) to wild-type ClpS. Wild-type Linker-GFP (blue) does not bind to ClpS.

We have identified an essential component of the N-end rule pathway in bacteria, the ClpA adaptor protein ClpS. This protein is directly responsible for the binding and delivery of substrates carrying the N-degron to ClpAP. Furthermore, we have shown that the N-degron in prokaryotes is more complex than previously understood. We propose that not only is a free α -amino group of the primary destabilizing residue required to initiate binding of ClpS, but an unstructured linker between the destabilizing motif and the rest of the protein is also a component of the N-degron required for effective transfer of the substrate from ClpS to ClpA. Positively charged secondary destabilizing residues and a positive net charge close to the N terminus are additional preferred features of the N-degron.

Most cellular proteins have stabilizing N termini. The low occurrence of N-end rule proteins is due to the specificity of the enzymes that modify N termini of proteins: namely, N-terminal acetylases and methionine aminopeptidases (MetAP)^{12–14}. *E. coli* MetAP, like its eukaryotic counterpart, shows a strong preference for small and uncharged residues in the penultimate position^{12,15}, thereby generating proteins with stabilizing N termini. The additional features of the N-degron required for N-end rule degradation (linker and charge distribution) further increase the selectivity of this pathway. The ClpAPS-dependent N-end rule pathway is thus ideally suited to regulate biological processes in bacteria. Consistent with a selective activation of N-end rule degradation are our preliminary findings

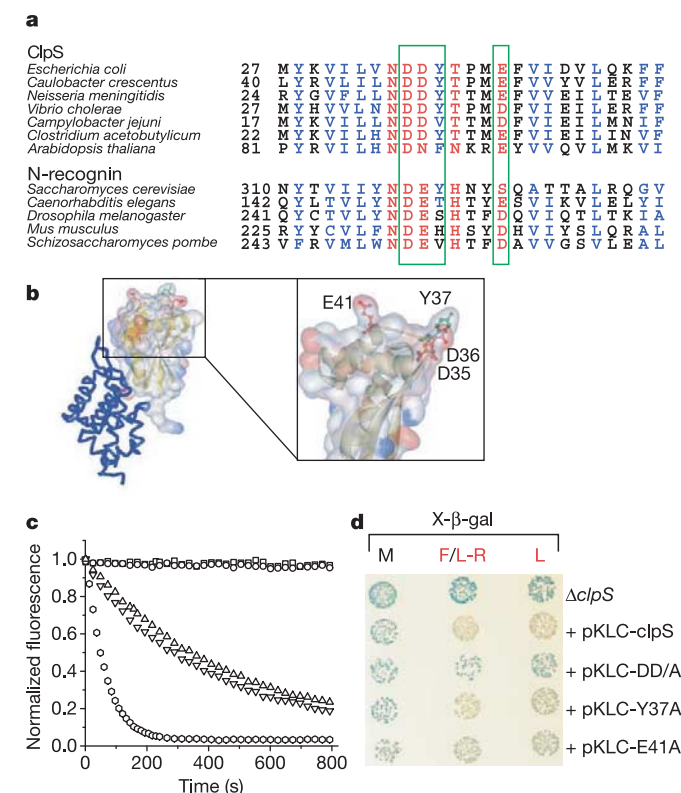


Figure 4 | ClpS has a specific binding site for N-end rule substrates. **a**, Sequence alignment of the N-end rule binding site of ClpS homologues and the putative binding site for type-2 N-end rule substrates of N-recognin homologues (blue, conserved hydrophobic residues; red, conserved charged residues; green rectangle, surface-exposed residues in ClpS). **b**, Structure of ClpS in complex with the N-terminal domain of ClpA⁴. Close-up shows surface-exposed conserved residues of the N-end rule binding site. **c**, **d**, Mutations in the binding site inhibit ClpS-mediated N-end rule degradation by ClpAP *in vitro* and *in vivo*. **c**, Degradation of FR-Linker-GFP by ClpAP mediated by wild-type ClpS (hexagons, lower curve), ClpS-Y37A (triangles), ClpS-E41A (inverted triangles) or ClpS-DD/AA (squares) or in the absence of ClpS (circles). **d**, Expression of ClpS-DD/AA cannot complement the Δ clpS mutant; ClpS-Y37A and ClpS-E41A show a similar phenotype to wild-type cells.

that, when grown in rich media, $\Delta clpS$ mutant cells show neither obvious phenotypes nor changes (as compared with wild type) in the protein pattern.

We consider that N-degrons may be generated by several mechanisms. Proteins may become selectively converted to N-end rule substrates by the regulated action of the F/L-transferase. It is intriguing that both pro- and eukaryotes are equipped with aminotransferases, which act to generate N-end rule termini. Alternatively, the N-degron may be dormant in the interior of a protein sequence and rendered accessible (that is, at the N terminus) only under specific conditions that result in endoproteolytic processing, autolysis or conformational changes. The importance of successive proteolytic events for generating N-end rule substrates is documented in eukaryotes: DIAP1, an apoptosis inhibitor in *Drosophila melanogaster*, is first cleaved by a caspase that liberates an intrinsic N-degron¹⁶; and SCC1/RAD21, a subunit of the cohesin complex, is first cleaved by separase ESP1 to free a masked N-degron in a process important to maintain chromosome stability¹⁷. For bacteria, however, the physiological N-end rule substrates remain elusive.

It is the adaptor ClpS that provides selectivity to the N-end rule pathway. Not only does ClpS have high specificity for N-degrons, it simultaneously inhibits the degradation of substrates by ClpAP that are not recognized by ClpS. This ensures that the activity of the ClpAPS machine is highly controlled and ideally suited for regulatory proteolysis. It is an intriguing and unexpected finding that, despite obvious differences between the N-end rule pathways of eukaryotes and prokaryotes, the first decisive step of substrate recognition is highly conserved between the kingdoms (Supplementary Fig. S1). Using similar substrate-binding pockets, ClpS targets N-end rule substrates to the ClpAP protease, whereas N-recognin targets N-end rule substrates for ubiquitination and degradation by the proteasome.

METHODS

Strains and plasmids. The $\Delta clpS$ mutant was constructed from wild-type *E. coli* (MC4100) as described⁶. We removed the selection cassette as described¹⁸. Plasmids pUB23-M, pUB23-R and pUB23-L encoding ubiquitin-M- β -gal, ubiquitin-R- β -gal and ubiquitin-L- β -gal, respectively, and pJT184 encoding Ubr1 were a gift from R. J. Dohmen. The low-copy plasmid pHKLC was derived from pHSG575 and carries the kanamycin selection cassette of pMPM-K3. For the *in vivo* complementation assay, cDNAs encoding wild-type ClpS, ClpS-Y37A, ClpS-E41A or ClpS-DD/AA were cloned into pHKLC.

Proteins. ClpA, ClpP, ClpS and GFP-ssrA were purified as described⁶. FR-GFP was constructed by QuickChange mutagenesis of the *gfp* gene. cDNA encoding FR-GFP or wild-type GFP was inserted into the vector pTYB11 (New England Biolabs) to create N-terminal CBP-intein fusion proteins. A DNA linker encoding the ten N-terminal residues of wild-type GFP was inserted by means of integrating long-primer PCR. FR-GFP, FR-Linker-GFP and Linker-GFP were expressed as intein fusion proteins and purified by one-step affinity chromatography according to the manufacturer's instruction (New England Biolabs).

In vivo half-live measurements. Plasmids encoding different ubiquitin-X- β -gal fusion proteins (where X is the engineered N-terminal residue of the β -gal mutant) were co-transformed with a plasmid encoding Ubp1, a yeast ubiquitin hydrolase, into wild-type *E. coli* or a $\Delta clpS$ mutant. Ubp1 cleaves ubiquitin from the fusion co-translationally and liberates X- β -gal. [³⁵S]methionine pulse-chase experiments to determine the half life of the different X- β -gal proteins were done as described³.

In vivo complementation assays. Wild-type *E. coli* (MC4100) and $\Delta clpS$ mutant cells containing pJT184, *lacIq* and pUB23-X were transformed with pHKLC containing the genes for wild-type *clpS*, *clpS*-DD/AA, *clpS*-Y37A or *clpS*-E41A. Overnight cultures were diluted to a relative absorbance at 600 nm of 0.0001 and 5 μ l was spotted on LB agar plates containing 500 μ M isopropylthiogalactoside and 100 μ g ml⁻¹ of 5-bromo-3-indolyl β -D-galactopyranoside (Sigma). Plates were incubated at 30 °C overnight.

Fluorescence. Fluorescence kinetic measurements were made on a LS50B luminescence spectrometer (Perkin-Elmer), with excitation at 400 nm and emission at 510 nm. The reactions were carried out in reaction buffer (50 mM HEPES, 0.3 M NaCl, 20 mM MgSO₄, 0.5 mM dithiothreitol and 10% glycerol; pH 7.5) containing an ATP-regenerating system (3 mM phosphoenol pyruvate,

20 μ g ml⁻¹ of pyruvate kinase and 5 mM ATP), 83 nM ClpA₆, 42 nM ClpP₁₄, 500 nM wild-type ClpS or ClpS mutants and the appropriate GFP substrate at 500 nM as indicated.

Peptide libraries. Peptides were attached C-terminally to cellulose membranes and prepared by automated spot synthesis¹⁹. The peptides were derived from the sequences of the N terminus of GFP or β -gal, or from the whole sequences of the λ CI DNA-binding domain, RepA and β -gal as described¹⁰. The libraries were composed of 13-mer peptides, which scanned the protein sequences with an overlap of ten residues between neighbouring spots. The peptide libraries were incubated with 500 nM ClpS in buffer A (10 mM Tris, 150 mM KCl, 20 mM MgCl₂, 5% sucrose and 0.005% Tween 20; pH 7.5). Detection and quantification of ClpS bound to the peptides was performed as described¹⁰.

Received 24 September; accepted 7 November 2005.

1. Varshavsky, A. The N-end rule: functions, mysteries, uses. *Proc. Natl Acad. Sci. USA* **93**, 12142–12149 (1996).
2. Varshavsky, A. The N-end rule and regulation of apoptosis. *Nature Cell Biol.* **5**, 373–376 (2003).
3. Tobias, J. W., Shrader, T. E., Rocap, G. & Varshavsky, A. The N-end rule in bacteria. *Science* **254**, 1374–1377 (1991).
4. Zeth, K. et al. Structural analysis of the adaptor protein ClpS in complex with the N-terminal domain of ClpA. *Nature Struct. Biol.* **9**, 906–911 (2002).
5. Baker, R. T. & Varshavsky, A. Inhibition of the N-end rule pathway in living cells. *Proc. Natl Acad. Sci. USA* **88**, 1090–1094 (1991).
6. Dougan, D. A., Reid, B. G., Horwich, A. L. & Bukau, B. ClpS, a substrate modulator of the ClpAP machine. *Mol. Cell* **9**, 673–683 (2002).
7. Lupas, A. N. & Koretke, K. K. Bioinformatic analysis of ClpS, a protein module involved in prokaryotic and eukaryotic protein degradation. *J. Struct. Biol.* **141**, 77–83 (2003).
8. Knoblach, N. T. et al. Substrate specificity of the SecB chaperone. *J. Biol. Chem.* **274**, 34219–34225 (1999).
9. Deuerling, E. et al. Trigger factor and DnaK possess overlapping substrate pools and binding specificities. *Mol. Microbiol.* **47**, 1317–1328 (2003).
10. Rudiger, S., Germeroth, L., Schneider-Mergener, J. & Bukau, B. Substrate specificity of the DnaK chaperone determined by screening cellulose-bound peptide libraries. *EMBO J.* **16**, 1501–1507 (1997).
11. Levy, F., Johnston, J. A. & Varshavsky, A. Analysis of a conditional degradation signal in yeast and mammalian cells. *Eur. J. Biochem.* **259**, 244–252 (1999).
12. Bradshaw, R. A., Brickey, W. W. & Walker, K. W. N-terminal processing: the methionine aminopeptidase and N α -acetyl transferase families. *Trends Biochem. Sci.* **23**, 263–267 (1998).
13. Hershko, A., Heller, H., Eytan, E., Kaklij, G. & Rose, I. A. Role of the α -amino group of protein in ubiquitin-mediated protein breakdown. *Proc. Natl Acad. Sci. USA* **81**, 7021–7025 (1984).
14. Mayer, A., Siegel, N. R., Schwartz, A. L. & Ciechanover, A. Degradation of proteins with acetylated amino termini by the ubiquitin system. *Science* **244**, 1480–1483 (1989).
15. Ben-Bassat, A. et al. Processing of the initiation methionine from proteins: properties of the *Escherichia coli* methionine aminopeptidase and its gene structure. *J. Bacteriol.* **169**, 751–757 (1987).
16. Ditzel, M. et al. Degradation of DIAP1 by the N-end rule pathway is essential for regulating apoptosis. *Nature Cell Biol.* **5**, 467–473 (2003).
17. Rao, H., Uhlmann, F., Nasmyth, K. & Varshavsky, A. Degradation of a cohesin subunit by the N-end rule pathway is essential for chromosome stability. *Nature* **410**, 955–959 (2001).
18. Datsenko, K. A. & Wanner, B. L. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl Acad. Sci. USA* **97**, 6640–6645 (2000).
19. Kramer, A. & Schneider-Mergener, J. Synthesis and screening of peptide libraries on continuous cellulose membrane supports. *Methods Mol. Biol.* **87**, 25–39 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. J. Dohmen for pUB23-M, pUB23-R, pUB23-L and pJT184; R. Volkmer for synthesis of peptide libraries; J. M. Weber for technical assistance; R. Nicolay for help with surface plasmon resonance; and laboratory members of the Clp group for discussions. This work was supported by the Deutsche Forschungsgemeinschaft priority program, proteolysis in prokaryotes: protein quality control and regulatory principles, the Fonds der Chemischen Industrie and the Australian Research Council QEII Fellowship to D.A.D.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.B. (bukau@zmbh.uni-heidelberg.de).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

Variations on a theme

Some say there are only seven or so basic types of plot, which get cast into myriad forms as we write and tell stories. The same might be true for the experiences of graduate students — the tales vary, but the underlying issues seem remarkably similar.

In our competition to find four new correspondents to write this year's Graduate Journal series, we received more than 170 entries from 24 countries. All were eager to document the highs and lows of their forthcoming year in the lab.

Despite the geographical spread and disciplinary diversity of the applicants — ranging from tiger genetics and cancer biology to synthetic chemistry and quantum information — the judging panel detected some common themes. The entrants were looking beyond graduate school to their professional future. They pondered staying in academic science or working in industry; pursuing traditional research paths or taking an alternative route. Many, especially women, worried about how to balance career and partner or family. Other concerns included whether or not to move abroad, and how to pick a good project.

Each entrant riffed on these themes in compelling ways,

to the extent that the judges felt they actually knew the people writing. The hopefuls were judged on story telling, insight, creativity and writing ability. It was a difficult choice, but we are pleased to reveal our four winners.

Milan de Vries, who writes his first column this week (see page 760), is a final-year molecular-biology student at the Massachusetts Institute of Technology.

Andreas Andersson is from Sweden and is a final-year oceanography student at the University of Hawaii. Katja Bargum is a final-year evolutionary-biology student at the University of Helsinki in Finland; and Mhairi Dupré is a first-year evolutionary-biology student at the University of Oxford, UK.

Congratulations to our winners — we look forward to following the twists and turns of their stories in 2006.



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Assistant Editor: Corie Lok

European Head Office, London
 The Macmillan Building,
 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

European Sales Manager:
 Andy Douglas (4975)
Natureevents: Silke Opstrup (4994)

UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
**Germany/Austria/
 The Netherlands:**
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

**European Satellite Office
 Germany/Austria/
 The Netherlands:**
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 75 Varick Street,
 9th Floor, New York,

NY 10013-1917
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

When personalities clash

Conflict in the laboratory can depress attitudes, dampen prospects or even derail young science careers. **Kendall Powell** finds out how to resolve issues and stay on track.

We've all heard horror stories about scientists behaving badly. The postdoc who threw pipettes around the laboratory. The adviser who forced lab members to look after her incontinent dog. And of course, the thesis committee member who made a pass at a visibly pregnant student.

Most conflict in the laboratory stems from common, less dramatic problems. Nevertheless, personality conflicts, poor communication and perceived favouritism can all make lab life miserable.

A difficult or uncomfortable relationship with an adviser can wreak havoc on a trainee's scientific career.

"I had slowly given up what I wanted to do without even knowing it," says Jacob*, a former molecular-biology graduate student at a US West Coast research institution. He was so burnt out by the strained relationship with his adviser that he gave up research after completing his PhD and took a job as a nature guide. Jacob eventually returned to an industry career and now accepts that his adviser taught him critical thinking and healthy scientific scepticism — even though it was painful at the time.

Workers in other labs agree that even the worst experiences taught them something about themselves and how to deal with sticky situations. But what can we learn from their experiences?

Friction

Jacob's conflict with his adviser arose from a scientific cultural gap. "He lived in the ivory tower," says Jacob. He was frustrated by his adviser's lack of empathy for a modern graduate student's concerns, such as the need to save money and the increased competition for permanent jobs. Jacob was looking for a career-building thesis project, but his adviser had other views: his boss's mantra was 'science for the sake of science'. These different expectations led to daily aggravation.

Differences in career goals, working hours and needs for publication are major sources of friction between principal investigators and their trainees.

Work hours are a personal choice, but if yours do not overlap with your supervisor's, tensions can build. "Reading research papers at home on a Saturday doesn't do you any good," says one former postdoc. Your adviser simply cannot see the effort you are putting in.

To avoid problems, be clear about goals and research milestones up front, and ask other lab

members what the boss expects when you join the lab.

Advisers sometime cross the line of acceptable behaviour by restricting holiday time or even forbidding workers to attend family functions that require long-distance travel. In these cases, it is important to set boundaries and explain that spending reasonable time on holiday or with your family will make you a more productive, less distracted scientist in the long run.

Clashes also occur frequently over publications. In frustration, one student submitted a paper without her adviser's permission. The move backfired and her angry adviser asked the journal to pull the article from review.

This is not an approach that one former departmental head at a UK university, Colin, would advocate. Instead, to trigger action on a manuscript, he advises: "Keep talking to the supervisor and suggest that you will write the experimental section." Try polite reminders first and get a third party involved only as a last resort.

Mary Bradley, director of postgraduate affairs at Washington University in St Louis, notes that many research universities have clear authorship guidelines, stating who should be included as an author and what contributions earn first author spots.

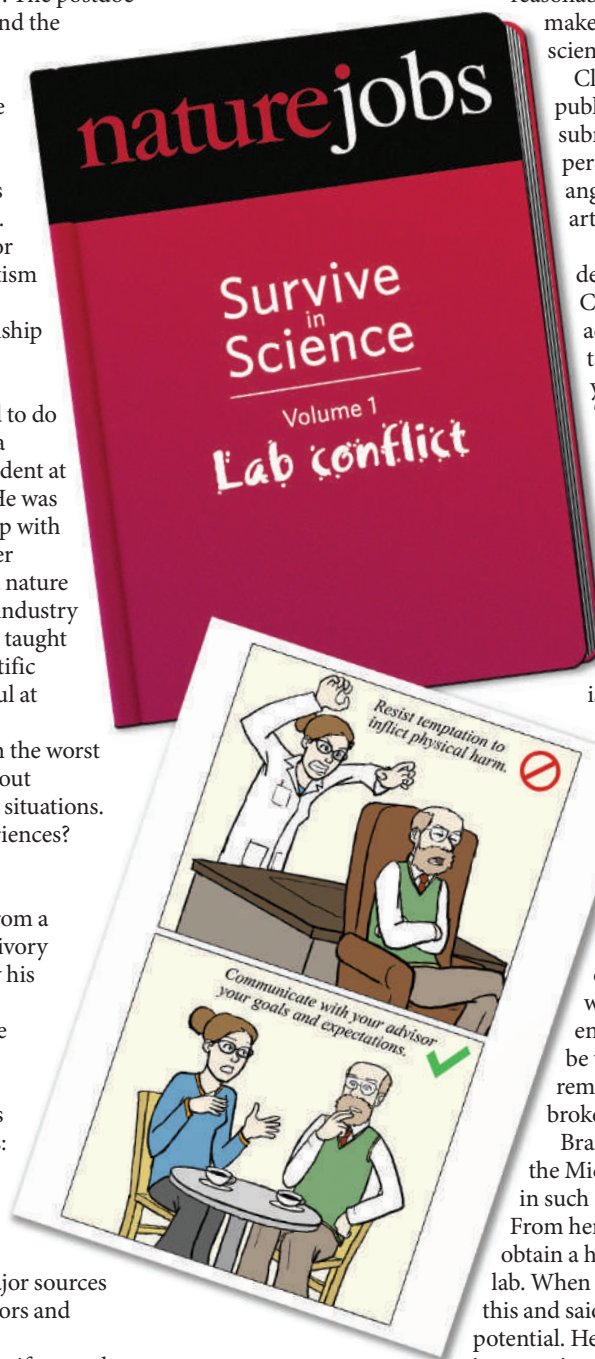
"Generally, it's a communication issue," says Bradley, who often counsels postdocs. "Make sure you fully discuss what you both think you are supposed to be working on."

Nightmares

But graduate students and postdocs who find themselves arguing constantly with an adviser, or worse, being ignored by their lab head, must make hard decisions. Is an interesting project worth continuing with in such a hostile environment? Would switching laboratories be too much of a career setback? Can you remain professional and productive with a broken mentor relationship?

Brandy, a former biology graduate student in the Midwest of the United States, found herself in such a relationship halfway through her thesis. From her perspective, she felt she was expected to obtain a higher standard than other students in the lab. When she confronted her adviser, he admitted this and said he felt she was not achieving her full potential. He criticized her writing, was unsupportive in committee meetings and asked her to do unnecessarily tedious or redundant experiments.

At first, she complained only to friends and family, because she expected things to get better. Then, her adviser asked her for a personal favour for his family. Brandy bristled at the idea of crossing a professional line and told him she was not comfortable doing the favour. After her refusal, her



Bad science: if you're unlucky enough to have a lazy or unhelpful supervisor, take advice before acting.

W. FERNANDES

J. CHAM

CORBIS

relationship with her adviser soured further.

"Knocking graduate students down and then trying to build them up to what you want them to be is not mentoring," Brandy says. She decided to stick it out because she did not think that complaining to others would help — her adviser had a big reputation at that university.

Instead, she cultivated a closer rapport with a thesis-committee member and made sure her science was infallible when it came to defending her thesis. She also learned an important lesson about herself — how not to be a doormat for others. "Now, I insist that people treat me with respect," she says.

Michelle, an environmental studies master's student based in the Midwest, also had a poor relationship with her adviser. She says he made unreasonable demands of her, such as asking her to drive ten hours for an appointment that turned out to be a ten-minute conversation, or making her learn obsolete computer software.

"I was more of a lackey to him than a student," she says. "At one point I had had enough and I basically quit." But after a weekend at home, she realized she wanted to graduate and returned to finish her project.

Michelle and Brandy survived because of the support they both received from other faculty members in their scientific field. "You need to have recommendations, if nothing else," says Michelle. "You have to prove to someone that you do good work and you are worth hiring."

Of course the adviser isn't always at fault. Colin put his own credibility on the line recently when he defended a student charged with unethical and illegal activity by the university. Believing the student's claim of a one-time offence, Colin lobbied the university not to end the career of a hard-working, productive and contrite student.

Resolutions

There are no short-cuts to conflict resolution, notes Colin. It usually requires large chunks of time to satisfy all parties. But many students and postdocs fear that they will be labelled as troublemakers if they turn to department or university officials for guidance. Also, trainees commonly perceive that faculty members will automatically stick together. But these fears may be unfounded, says Colin.

"Students should have more confidence that other academics will take a detached view and treat both parties as equals. In general, colleagues do not protect an academic if he doesn't deserve protecting," he says.

Bradley says most postdocs just want someone to listen to their grievances. As an advocate for postdocs, she puts people in touch with the correct university offices for serious concerns such as research fraud or sexual harassment. For dealing with personality clashes, she gives pointers on how to overcome the "dreaded fear of having a difficult conversation with their adviser".

Her advice? Do it sooner rather than later — procrastinating will only result in more stress. Use 'I' statements and not accusatory 'you' statements, as in,

IMAGE UNAVAILABLE FOR COPYRIGHT REASONS

"I was under the impression that I would be first author on this paper, but then John became first author and that really frustrated me." Make a list of three to five things you want to discuss.

"This is your career. You are going to have to talk about it. Then let the investigator's response determine what you do next," says Bradley. A more formal route might include a visit to the university ombudsman's office to investigate mediation.

"We offer coaching and role-playing to help people acquire the confidence to work things out on their own," says Lois Price Spratlen, the university ombudsman at the University of Washington in Seattle. After a confidential meeting with Price Spratlen, and if a client does not want to proceed alone, a mediation procedure is set into motion.

From Price Spratlen's office, the client calls the manager of the person about whom there is a complaint. That manager — an investigator, department chair or dean — has 24 hours to get the problematic person to make an appointment with Price Spratlen. After she meets that person in private, she brings both parties and the supervisor to the mediation table. Her job is not to take sides, but to get the two sides talking.

Not everyone wishes to go forward with mediation. "When people are uncertain what to do, I give them my 24-hour rule," says Price Spratlen. "Take 24 hours to think about ways you have managed conflict before and whether those methods would help in this situation," she says.

She also advises clients to write down their story. Finally, she relies on the golden rule — treat others as you wish to be treated — for keeping emotions in check.

Perhaps a bit of conflict and suffering is necessary in every young scientist's life, but communicating your own needs and expectations to your adviser can make a miserable situation more tolerable.

"You won't know what is a positive or negative event in your life until the end," reflects Jacob. He even questions whether he should have ever applied to graduate school. But in the end, he says, it has worked out well. "Now I have this job, which is what I really want to do with my life."

Kendall Powell is a freelance science writer in Broomfield, Colorado.

*Some names have been changed to protect privacy.



Mary Bradley: don't procrastinate, communicate.

LENERO

MOVERS

Hendrik Weerts, director, high-energy physics division, Argonne National Laboratory, Argonne, Illinois



2004–05: Sabbatical at Fermi National Accelerator Laboratory, Batavia, Illinois
1998–2002: Visiting scientist, Fermi National Accelerator Laboratory, Batavia, Illinois
1983–2005: Professor, Michigan State University, East Lansing, Michigan

As a high-school student, Harry Weerts finished reading his physics textbook, but was left with more questions than answers. His quest to find those answers led him from his native Netherlands to the forefront of particle physics in the United States.

From the start, Weerts had no doubt that he would study physics at university. But it was a riveting class about neutrino scattering at the RWTH Aachen University in Germany that compelled him to do both his masters and his PhD at the same university, working on neutrino experiments at CERN, the European particle-physics lab near Geneva, Switzerland.

His biggest career decision, he says, was to move to the United States and work at the Tevatron — the world's highest-energy particle accelerator — as a postdoc at the Fermi National Accelerator Laboratory (FermiLab) in Batavia, Illinois.

From there he took a faculty position at Michigan State University in 1983, where he was able to spend the next 20 years working on one of the two large-collider experiments designed at the Tevatron. He oversaw the programme — which continues to collect data today — since it began running in the 1990s.

A self-described “lucky guy”, Weerts was part of the team working at the Tevatron that in 1995 discovered the largest elemental subatomic particle: the elusive top quark.

Always on the lookout for the next big collaborative project, Weerts recently took a year's sabbatical at FermiLab to champion the International Linear Collider — the proposed particle accelerator that would be able to create high-energy collisions between electrons and positrons.

Most recently, he decided to put off retirement and give up his tenure at Michigan State to serve as director of high-energy physics at Argonne National Lab in Argonne, Illinois. This decision was driven by his belief that the younger cohort should conduct the experiments and that it is the duty of the more senior scientists to lead their field in new directions.

“I feel I owe it to my research field to help define the next big project in particle physics,” Weerts says. Being a director at Argonne enables him to make this contribution at an administrative level, he adds. And with so much left to discover, he wasn't quite ready for retirement just yet.

“Only 4% of the Universe is made out of the stuff we've studied for the past 2,000 years,” he says. ■
Virginia Gewin

SCIENTISTS & SOCIETY

Dabbling in science journalism

Somewhere between working on my research project, panicking about deadlines and teaching undergraduate students, I took a graduate course in science communication at the Karolinska Institute in Stockholm, Sweden. The rationale behind it is that scientists must accept responsibility for communicating their results to a wide audience.

The course — held one evening a week this past autumn and taught by journalists, government administrators and scientists — covered many aspects of science journalism, ranging from ethics, graphics, broadcast and magazine reporting to the overall portrayal of science in the media.

As part of the course, I did a two-week internship at the *Nature* office in Washington DC with *Naturejobs* editor Paul Smaglik. He showed me how he runs his sections and let me help with some of the editing tasks. I discovered the difficulties in shortening and altering other people's work. I talked to some other *Nature* staff, and asked the manuscript editors about the editing and peer-review process.

I also found time for reflection. I learned to appreciate how significant communicating about science and medicine is to the education of the general public. Journalists are key

players in this process, acting as interpreters between scientists and the public. During this interpretation, though, messages can become distorted and that can strongly influence public opinion. Perhaps we scientists are partly to blame: we spend years learning new laboratory techniques but spend far less time thinking about how best to deliver our newly found knowledge to the public.

More importantly, in a democratic society, controversial issues in research should be open and accessible to the public, which has a natural interest in knowing how taxpayers' money is spent. Scientists should play a key role in shining a light on these issues, and one way to do that is to engage with the media. It is also in the individual researcher's own interest to do this, to generate interest, gain recognition and explain the need for funding.

This course taught me that the public perception of science has become almost more important than the science itself. I hereby promise that in my career, I will keep a straight and honest relationship with medical journalists. If that turns out to be too difficult, well, then I might just have to become one myself. ■

Lina Nordquist is a PhD student at Uppsala University in Sweden.

GRADUATE JOURNAL

PhD TV

They'll never make a TV comedy about graduate students. I know this because I just spent three days watching all of the second season of *Scrubs* — the goofy American comedy about would-be doctors doing their residency. *Scrubs* has three things that a show about graduate students wouldn't.

One, people actually die on *Scrubs*. “Professor Frank, I just don't think this *E. coli* is going to pull through” is unlikely to drive ratings.

Two, they have patients. The best we can offer are ideas and blips on a graph. Bloody operations are more fun to watch than percolating ideas.

Three, people have been to see doctors. They might reasonably wonder what it's like to be one. By contrast, tell someone at a party you're in science and the next question tends to be “So, how's the punch?” As grouchy Dr Cox would summarize on *Scrubs*: “Viewers just don't care about graduate students, Bambi.”

This really is a shame. Consider me: I'm in my last year of a PhD, looking for a postdoc in a field that's fundable, not super-competitive, yet super-hot for 15 more years. I keep daydreaming about careers that society respects more, like plumbing. My mom keeps calling to tell me about my friends who have bought houses. I spend more time talking to yeast than to people. My doctor tells me I stress too much. Oh, and I need to graduate. Soon. How will all this unfold? Stay tuned. If TV executives happen to be reading this, call any time. I'll be in lab. ■

Milan de Vries is a molecular-biology graduate student at the Massachusetts Institute of Technology in Cambridge, Massachusetts.

Photons do not lie

The true value of history.

Euan Nisbet

The camera sweeps down the majestic garden approach, the marvellous avenue of intertwined rose-chrysanthemum trees. Cranes flock above. Skittering impalas graze amid bamboo-gums, in which koalas and lumbering pandas loll under the warm sun. Onwards we glide to the palace and into the grand throne room, thronged with a vast crowd, the notables of Earth. They rise. In Reykjavik, the Nobel ceremony has begun. Entering in state to the ancient imperial anthem, *Rock Around the Clock*, is Her Majesty, Queen Hillary IV.

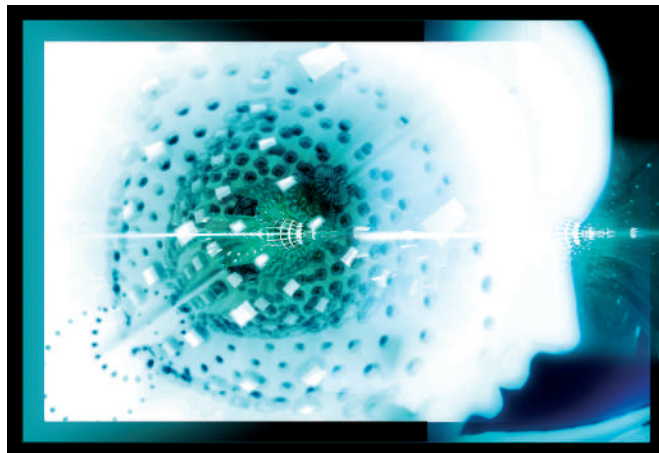
The laureates are led in. What paragons they are! There had been criticism that the biology prize should be discontinued, as biology is so unchallenging. That was surely correct for stamp-collecting physics, but who cannot applaud this year's biology winner, recreator of the rabbit-sized, slipper-and-iPub-fetching house-trained woolly mammoth? No maths prize, after that laureate's affair with the Prince Consort. How dare she! Of course, the discovery of the soul led to calls for reinstatement, but the Gödel-Ishōdad of Merv-Epimenides theory shows that, although detectable, that 'brane cannot be investigated, frangible yet intangible. The quest is abandoned.

The heavyweight prizewinners come on. Geology — for the discovery of the tenth inhabited planet, this one orbiting Aldebaran. The inhabitants are aggressive ant-like beings in the throes of nuclear war, but no doubt they will see reason when we pacify them. Such a pity that we had to sterilize all the humanoids of the eighth planet: the peace prize goes to the Admiral of that expedition, for those eloquent words of consolation, after the ritual welcoming dance offered by the inhabitants accidentally unveiled one of the Admiral's men. The strong reaction was of course fully justified.

Next is the economics prize. As is customary, this goes to the ranking associate professor at Lunar Chicago University, whose 'great idea' was the reinvention of crime. It is now proven that economic systems are unstable unless criminals flourish. Old fogies decry this, but the huge 'Make

Crime Real' demo outside the palace has much sympathy. Several provinces now select lucky teenagers for training, providing necessities such as body piercings, drugs and 'attitude'.

A hum of excitement pervades the room. It is time for the Sacred Reminder. The prime minister rises to recount the saving of humanity. Suddenly, rudely, there is a slight buzz. A masculinist suf-



frager has managed to enter. He is quickly ejected, burka fluttering. The prime minister dryly observes that the formula limiting men to 10% of academic posts is already over-generous.

The Reminder begins: the sterilization of Earth in 2020 by the week-long Great Solar Flare. She recounts the wondrous chance that the US president was on lunar darkside, in Lunar City to celebrate the close of her third term, accompanied by the Chinese president, the British Crown Prince and the Crown Princess of Japan. The presidents' energy and constitutional wisdom inspired extraordinary efforts to make the colony viable, and then rebuild. Then came the romance between Prince and Princess, and the final emotional abdication of the Dual Presidency, with constitutional dispensation of our Empire of Luna and Three Planets, into the hands of their beloved adopted heirs.

Since then, what progress there has been under the Rose and Chrysanthemum! From bones and buried seeds we have recovered the biosphere, even raising Earth's population to more than a million. Mirrored Venus was terraformed; Mars warmed by chlorofluorocarbons, and methane from polar peat bogs. The end of physics came with the discovery of CUTE, the canonic universal theory of everything, bringing instantaneous

travel to anywhere in the cosmos.

Last comes the greatest prize. What is more valuable than history, recovery of our lost past? The flare frizzled the databases, leaving only tiny fragments of our culture in the personal oddments of scientists in Lunar City: the deficient aesthetic of Bach, Mozart, Shakespeare, Milton, Keats — Jane Austen, for goodness' sake. But now we unearth a marvellous treasure trove. Of

course even the geeks loved Elvis and Abba, but only now do we fully realize just how limited was their vision, they who did not even value the great Sinatran song of creation: "Do BE! Do BE! DO!"

Now we are recovering the true genius of the Golden Age. Here is our history prizewinner now, rising to deliver her lecture. She recounts the long hours spent cruising round the wavefront, laboriously integrating photon after photon on the sphere 200 light years out, as she collected the

radio broadcasts of the twentieth century.

What diamonds she has found! One is the 00.48 a.m. BBC shipping forecast, a mysterious sacred compline into the dead of night, with its great unsung hymn of 'Sailing by' into the afterlife. From American radio, the account by Orson Welles of the 1938 martian attack on Earth is previously known only in a corrupt fake purporting to have been written by a Mr H. G. Wells. Already major funding supports archaeologists searching Mars for the extinct civilizations under our new oceans.

Finally comes the climax of her Nobel lecture. Technically, it is far easier to recover radio than TV. But now, a visual revelation — proof that twentieth century humanity was in touch with alien life elsewhere in the Universe! Praise be: these aliens, delegated to each nation but unknown to us moderns, are fully humanoid, female of course, possessing unearthly beauty and uttering truly sublime speeches for World Peace. Here is the highest glory of the Age of Gold. Here from civilization's ancient heartland — Perth, Western Australia — gentlemen and noble ladies, I give you the Miss Universe competition, 1979!

Euan Nisbet puzzles over Archaean rocks and monitors atmospheric greenhouse gases at Royal Holloway, University of London, Egham, UK.

JACEY